HELSINGIN YLIOPISTO

# Suitability of Neural Machine Translation for Different Types of Texts

## A study on potential predictors

Ari Gröhn
Pro Gradu Thesis
English Translation
Department of Modern Languages
University of Helsinki
April 2019

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

| Tiedekunta – Fakultet – Faculty | Laitos – Institution – Department | |
|---|---|---|
| Humanistinen tiedekunta | Nykykielten laitos | |
| Tekijä – Författare – Author | | |
| Ari Gröhn | | |
| Työn nimi – Arbetets titel – Title | | |
| Suitability of Neural Machine Translation for Different Types of Texts: A Study on Potential Predictors | | |
| Oppiaine – Läroämne – Subject | | |
| Englannin kääntäminen | | |
| Työn laji – Arbetets art – Level | Aika – Datum – Month and year | Sivumäärä – Sidoantal – Number of pages |
| Pro Gradu -tutkielma | Huhtikuu 2019 | 47 s., suomenkielinen lyhennelmä 10 s. |

Tiivistelmä – Referat – Abstract

Tutkielmassa tarkastellaan erilaisten tekstien soveltuvuutta neuroverkkokonekääntämiselle. Tutkimus pyrkii löytämään kielellisiä indikaattoreita, joita voidaan käyttää ennustamaan, onko jokin tietty teksti soveltuva neuroverkkokonekääntämiselle vai ei. Koska aihetta ei ole vielä tutkittu laajasti, tutkimuksessa esitetään myös erilaisia tutkimustapoja, joilla aihetta voisi tutkia.

Tutkielman teoriatausta muodostuu tekstityyppien tutkimuksesta ja neuroverkkokonekääntämisestä. Lähdekirjallisuuden perusteella soveltuvimmaksi tekstityyppiluokitteluksi nousee Biberin viisi dimensiota, joita käytetään materiaalivalinnassa ja joiden yhteyksiä käännöslaadun kanssa tarkastellaan analyysin aikana. Neuroverkkokonekääntämisen osalta esitellään lyhyesti neuroverkkokääntimien eroavaisuuksia aiempiin kääntimiin, neuroverkkokäänninten perusrakennetta sekä niille tyypillisesti vaikeita kielellisiä elementtejä.

Tutkielmassa käytetään materiaalina kolmea eri korpusta, jotka ovat fiktio, viralliset kirjeet ja viralliset dokumentit. Kukin korpus koostuu alkuperäisestä englanninkielisestä lähtötekstistä, suomenkielisestä ihmisen tekemästä referenssikäännöksestä sekä kahden neuroverkkokonekääntimen käännöksestä. Korpukset analysoidaan automaattisella evaluaatiolla ja kustakin korpuksesta otetaan pienempi otos, jolle tehdään manuaalinen virhekategorisointi. Näin tutkimus vertaa erityyppisten tekstien konekäännösten laatua toisiinsa ja tutkii, onko käännöksissä tapahtuneiden virheiden välillä merkittäviä eroja erilaisten tekstien sekä kahden kääntimen välillä. Tekstityyppien lisäksi tutkimuksessa tarkastellaan lausepituuden suhdetta käännöslaatuun, joka on yksi lähdekirjallisuudessa havaituista käännöslaatuun vaikuttavista tekstuaalisista piirteistä.

Tutkielmassa käytettyjen kolmen korpuksen perusteella selviää, että Biberin dimensioista narratiiviset tekstit näyttäisivät olevan huonommin soveltuvia neuroverkkokonekääntämiselle kuin ei-narratiiviset ja että kontekstisidonnaiset tekstit olisivat huonommin soveltuvia kuin eksplisiittiset. Fiktiokorpuksen virhejakauma eroaa eniten kahden muun tuloksista, mutta tutkielmassa käytetty materiaali havaitaan mahdollisesti ongelmalliseksi. Konekäänninten välillä havaitaan joitain eroja, mutta niiden syitä on vaikea arvioida tuntematta tarkemmin kääntimien rakenteita. Lausepituusanalyysin perusteella lyhyempiä lauseita voidaan käyttää yhden korpuksen sisällä ennustamaan tulosta, mutta korpusten välinen vertailu ei ole mahdollista ja äärimmäisen lyhyet lauseet saattavat olla muista syistä ongelmallisia.

Analyysin perusteella päätellään, että Biberin tapaista kielellisiin piirteisiin perustuvaa tekstityyppiluokitusta voidaan jossain määrin käyttää ennustamaan erilaisten tekstien soveltuvuutta neuroverkkokonekääntämiselle, joskin lisätutkimusta vaadittaisiin asian kattavaan kartoitukseen. Tutkimuksessa käytetyt menetelmät havaitaan pääasiassa hyviksi asian tutkimiselle, joskin virheluokitteluun esitetään pientä tarkennusta.

Avainsanat – Nyckelord – Keywords
neuroverkkokonekääntäminen, konekääntäminen, kääntäminen, tekstityyppi, käännöslaatu

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information

# Index

1 Introduction ..... 1
2 Theory ..... 3
2.1 Text types ..... 3
2.1.1 Qualitative text type categorisations ..... 4
2.1.2 Quantitative text type categorisations ..... 6
2.1.3 Discussion ..... 9
2.2 Neural Machine Translation ..... 10
2.2.1 The basics ..... 11
2.2.2 Characteristics of NMT output ..... 13
2.2.3 Discussion ..... 15
3 Material and method ..... 17
3.1 Material ..... 17
3.1.1 Covering the five dimensions ..... 17
3.1.2 Material used in the study ..... 19
3.2 Methods ..... 20
3.2.1 Error typology for sentence-level analysis ..... 21
3.2.2 Automatic evaluation ..... 22
3.3 NMT engines used in the study ..... 24
4 Analysis ..... 26
4.1 Corpus level analysis ..... 26
4.1.1 LeBLEU scores ..... 26
4.1.2 Back to the dimensions ..... 28
4.2 Segment level analysis ..... 29
4.2.1 Frequent errors per corpus ..... 31
4.2.2 Comparing the two engines ..... 35
4.3 Sentence length analysis ..... 36
5 Conclusions ..... 39
References ..... 44
Appendix 1: Lyhennelmä ..... 1

# List of figures

Figure 1: Mean scores of dimension 1 ('Involved informational production') for nine genres (figure from Biber 1989: 12) ..... 8
Figure 2: Error type frequencies for the Fic corpus ..... 32
Figure 3: Error type frequencies for the PL corpus ..... 32
Figure 4: Error type frequencies for the OD corpus ..... 32
Figure 5: Equally weighted error distribution between Lingsoft's and Google's NMT engines ..... 35

# List of tables

Table 1: Biber's five dimensions and their linguistic features (Biber 1989: 8-9) ..... 7
Table 2: Simplified placing of three subgenres in each of the five dimensions and their scores ..... 18
Table 3: Corpus level LeBLEU scores ..... 26
Table 4: Evaluation scores of the three genres in Biber's dimensions ..... 29
Table 5: Sentence length in all material vs. LeBLEU score ..... 37
Table 6: Relation of LeBLEU score to average sentence length ..... 37

# List of examples

Example 1: Professional letter in the material ..... 19
Example 2: Official document in the material ..... 20
Example 3: Fiction in the material ..... 20
Example 4: Creativity in the Fic corpus ..... 27
Example 5: Segment in the PL corpus ..... 27
Example 6: Valid 0.000000 score ..... 30
Example 7: Semantic error in segment with high score ..... 30
Example 8: Semantic error ..... 31
Example 9: Omission in the OD corpus ..... 34
Example 10: Unnecessary whitespaces in Google's NMT output ..... 35
Example 11: Google fails to parse together entire sentence ..... 36
Example 12: Short segment with 0.000000 score ..... 37
Example 13: Creative liberties in short segment ..... 38

# List of abbreviations

| | |
|---|---|
| ALPAC | Automatic Language Processing Advisory Committee |
| BLEU | Bi-Lingual Evaluation Understudy (evaluation metric) |
| Fic | fiction (corpus in the study) |
| LeBLEU | Letter-Edit-BLEU/Levenshtein-BLEU (evaluation metric) |
| MT | machine translation |
| NMT | neural machine translation |
| OD | official documents (corpus in the study) |
| PBSMT | phrase-based statistical machine translation |
| PL | professional letters (corpus in the study) |
| RBMT | rule-based machine translation |
| SMT | statistical machine translation |
| TM | translation memory |

# 1 Introduction

Machine translation (MT) has been around since the 1940s and several iterations, notably rule-based MT (RBMT) and corpus-based MT (SMT), were developed before the turn of the century (Hutchins 2007). The era of widespread, commercial neural machine translation (NMT) kicked off in 2016 when, among others, Google announced its new NMT engine in 2016 (Wu et al. 2016). NMT differs from previous iterations by utilising neural networks and deep learning (Forcada 2017). This has been shown to greatly increase text fluency, while some errors, such as those related to accuracy, are still prevalent (see chapter 2.2.2).

The purpose of this thesis is to discover whether there are linguistic features that can be used to identify whether a particular text might be suitable or unsuitable for neural machine translation (NMT). The significance of potential results should be obvious, as they could allow those working with translations to identify whether the translation process of a given text might benefit from being machine translated and post-edited instead of being translated in full, and to eventually predict post-editing costs. Post-editing has been shown to increase translator productivity (with specifically NMT output increasing it even more than SMT output, see e.g. Shterionov et al. 2018). My hypothesis is that the quality of the machine translation will vary, and could be predicted, based on the text in question.

My research questions are, thus: What are the most suitable texts for neural machine translation? Are there any linguistic features that can be used to predict whether a text is suitable for machine translation? What are these linguistic features? And how can these features be studied? This study will look at these questions using automatic evaluation as well as some manual error categorisation. The study is not interested in the *quality* of the NMT output per se, but how its output differs from translations carried out by human translators of the same texts, following the leading principle of automatic evaluation: "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al. 2002: 1).

The subject is one with relatively little research and, as far as I am aware, no one has yet carried out broader research on how NMT specifically impacts the suitability of different texts for MT. The closest is a study by Calude (2003) that dates back 15 years, to a time when machine translation was far from today's quality, and that was carried out using manual evaluation of the texts. Another, more recent study by Salimi (2014) looks at the accuracy of Google Translate with fictional and non-fictional texts

but, as a bachelor's degree project, the scope of the study is limited and it uses the previous, SMT-based model of Google Translate instead of the newer commercial NMT engine.

The study progressed as follows: First I identified a suitable text type categorisation by Biber (1988, 1989) to use as a basis for choosing a selection of texts for the analysis. Three genres were identified in Biber's text type categorisation, or *dimensions*, that were suitable for the study: fiction, professional letters and official documents. A selection of these texts was then gathered for three individual corpora and translated using two separate NMT engines. Next, a quantitative analysis was carried out using automatic evaluation to compare the three corpora to see if and which Biber's dimensions appeared to show correlation. A segment level, qualitative analysis was also carried out on 180 individual segments in the three genres to see which types of errors they appeared to make and whether the three genres and two engines appeared to make different types of errors. For the error type categorisation, the DQF-MQM error typology by TAUS (2019) was used. Finally, a sentence length analysis was carried out to see whether there was correlation between the average score and sentence length of the genres.

The main theoretical concepts of this thesis are: neural machine translation (NMT), text types and automatic evaluation. The theory will focus on text types rather than genres as text types are directly related to the *linguistic* aspects of texts, whereas genres are mostly in relation to *contextual* aspects (see chapter 2.1). Biber's dimensions were, however, used as a basis for choosing the three genres in the study. This thesis is only interested in NMT from a linguistic perspective and how it can be utilised, instead of focusing on the technical side of how it functions. As such, the focus will be on comparing NMT to previous iterations of MT and considering how that might show in its output. The thesis will also briefly discuss automatic evaluation and why LeBLEU was chosen for the study over more established options, namely BLEU.

The rest of this study is organised as follows: First, chapter 2 introduces two key theoretical concepts: text types and NMT in chapters 2.1 and 2.2, respectively. Chapter 3 discusses material selection and the methods used in its analysis in chapters 3.1 and 3.2. Chapter 4 describes the analysis and its results. Finally, chapter 5 draws together what was discussed previously and evaluates the study from a broader perspective.

# 2 Theory

This chapter will focus on text types and neural machine translation (NMT). Chapter 2.1 discusses different text type categorisations, starting from qualitative text types in chapter 2.1.1, moving to qualitative in 2.1.2 and evaluating their suitability for the study in 2.1.3. Chapter 2.2 focuses on NMT, giving a basic overview of it in chapter 2.2.1, presenting the error types that appear to be most prevalent for it compared to previous iterations of MT in 2.2.2 and discussing how things relate to this thesis in chapter 2.2.3.

## 2.1 Text types

The question of what a text type is has many answers. Authors and researchers have used the term for different purposes over the years and there is no single dominant definition. Some, such as Reiss (1977), use text type to describe the functionality of a text, whereas some, like Werlich (1976), use it to define contextual foci. Reiss's and Werlich's text type categorisations could be seen as *qualitative*, or that they have looked at the meanings or functions of different texts and attempted to find features that link some of them together. There are also *quantitative* methods of text type categorisation, most notably by Biber (1988, further refined in 1989). Quantitative text type categorisations are based on corpus analyses where a large amount of text has been run through a program which has been told to identify clusters of texts based on pre-defined linguistics features.

The term *text type* is also closely related to *genre*—and sometimes used interchangeably—while some authors have also preferred to use a variety of terms such as *register*, *discourse* and *style* to describe what is largely the same phenomenon (see Shore & Mäntynen 2006: 38-39 and Lee 2001 for specific examples). Genres and text types do overlap to some extent and different text type categorisations could, for instance, be used to describe the differences between some genres (Shore & Mäntynen 2006: 37). The major difference is that a text representing a single genre can feature multiple different text types (ibid.). A newspaper article (a genre), for example, will often include descriptive, narrative and expository parts (text types, at least according to Werlich 1976) (ibid.). Genres (and on a lower level, texts) can, thus, be seen to *represent* certain text types but the two are still separate concepts. It should also be noted that this chapter discusses the matter from a purely linguistic perspective, even though the term *genre* has been used in various other fields as well: namely in literary, art and media studies (Shore & Mäntynen 2006: 19).

The ambiguity behind the concept of text type might lie in the differences in how the term has been used by German linguists compared to linguists in English-speaking countries. In German translation studies, a separation is made between text type (*Texttyp*) and text class[1] (*Textsorte*). Text type is used as a *functional* classification (e.g. informative vs. expressive), whereas text class as a *contextual* classification, i.e. where the text has been published (e.g. newspaper article, book). English-speaking linguists, however, sometimes consider the concept of text type to encompass both *Texttyp* and *Textsorte*. (Nord 2005: 20.)

The same separation can be seen in Finnish linguistics, supposedly due to the influence of German linguists. Text type (*tekstityyppi*) is generally used only to refer to Werlich's text types whereas text class (*tekstilaji*, *genre*) when describing the contextual differences (see e.g. Shore & Mäntynen 2006, Pietikäinen & Mäntynen 2009). Notably, Finnish linguists use *genre* (genre) and *tekstilaji* (text class) interchangeably (see Heikkinen et al. 2012).

For the purposes of this thesis, *text type* will be used when discussing the linguistic aspects of a text and *genre* to describe contextual aspects. Next, two of the most prevalent qualitative text type categorisations by Egon Werlich (1976) and Katharina Reiss (1977) are introduced in chapter 2.1.1. Then, it is discussed how corpus analysis uses linguistic markers to identify text types quantitatively and introduce Biber's dimensions (1988, 1989) in 2.1.2. In chapter 2.1.3, the chapter is drawn together from the perspective of the thesis.

### 2.1.1 Qualitative text type categorisations

A notable text type categorisation comes from Egon Werlich (1976) who categorises texts into five distinct categories based on what he calls their dominant contextual focus. Werlich acknowledges that the concept of text type is always "an idealised norm of distinctive text structuring" and uses the following five categories as examples of strategies that the one performing the illocutionary act can choose (1976: 39-41):

1. Descriptive
2. Narrative
3. Expository
4. Argumentative
5. Instructive

[1] *Textsorte* has also been translated as *text variety* by Chesterman (in Reiss 1977).

**Descriptive texts** are related to observable phenomena and marked by the use of adjectives and verbs related to existing or observing. **Narrative texts** look at different phenomena and how they relate to time. Their linguistic features include descriptive and dynamic verbs as well as often the use of past tense. **Expository texts** describe abstract ideas and their relations. They are often in present or present perfect tense and feature phrases denoting the relations between different phenomena. **Argumentative texts** consist of language denoting the differences between phenomena using, for instance, certain conjunctions (*but, though*), adjectives (*right*, *wrong*) and nouns. **Instructive texts** seek to direct, oblige or order someone to do something. This is achieved using, for instance, imperatives. (Werlich 1976: 39-41; Shore & Mäntynen 2006: 36-37.)

Katharina Reiss's (1977; Munday 2013: 73-74) text type categorisation, on the other hand, is interested in the communicative purpose of the text—or, in other words, its function—and looks at texts from the perspective of a translator. To Reiss, identifying a source text's text type is imperative to understanding which translation strategies should be used and which elements of the text should be prioritised and saved during the translation process (Reiss 1971; Fawcett 1997: 104). The categorisation is based on Bühler's (1934) three language functions for: *Darstellungsfunktion* (informative function), *Ausdrucksfunktion* (expressive function) and *Appellfunktion* (appellative function). Reiss's three[2] text types are:

1. Informative
2. Expressive
3. Operative

The goal of **informative texts** is the "plain communication of facts" (Reiss 1977: 108). They are usually logical or referential and focused on the content being transmitted. **Expressive texts** are a "creative composition" (ibid.) that focus on form instead. If language is just a tool in informative texts, in expressive texts it creates a substantial part of the text's value and brings the author to the front. No pre-defined content is being transmitted and the author chooses which thoughts to convey. **Operative texts** seek to induce "behavioural responses" (ibid.: 109) in the reader by appealing to them through persuasion. (Reiss 1977: 108-109; Munday 2013: 72.)

[2] Reiss's original classification also includes a fourth text type: audiomedial. The fourth type is not a separate one but encompasses all the other three. Reiss sees audiomediality as an additional circumstance that "needs special attention" in the translation process. (Reiss 1977: 111.) As such, it is irrelevant for the broader discussion on text types in this thesis.

Reiss's text types are not mutually exclusive, and most texts are a mixture of all three. A poem might be purely expressive, but a play or satire is expected to provide commentary on contemporary issues, thus making them both informative and operative. (Reiss 1977; Munday 2013: 72-73.) According to Reiss, this is not an issue as there should always be one text type that dominates over the others (Reiss 1977; Fawcett 1997: 104).

To summarise, Werlich and Reiss look at texts from slightly different angles. Werlich's text types are strategies for the one performing the illocutionary act and attempts to classify texts based on their contextual focus, whereas Reiss uses her text types to identify the key features of a text so that a translator knows which parts of a text to prioritise to convey its function properly. Werlich's text types could naturally be used for the same purposes as Reiss's, but they also include a linguistic aspect and seek to identify features that occur often in the texts. Next, quantitative text type categorisations by corpus linguists are discussed.

### 2.1.2 Quantitative text type categorisations

The other large group of text type categorisation is based on corpus analyses. Instead of sifting through text masses manually to categorise them based on their function and/or meaning, corpus linguists have sought to automate the process with the help of computers and a quantitative method of analysis. This has the obvious benefit of being able to analyse much larger amounts of data and, as such, reach broader conclusions. Automatic text categorisation has, thus, been a topic of a great deal of research over the years. Linguistic features have been used to automatically identify, for instance, genres (see e.g. Karlgren & Cutting 1994; Kessler et al. 1997) and authors (Stamatatos et al. 2000; Homem & Carvalho 2011).

The best-known example of quantitative text type categorisation comes from Biber (1988, further refined in 1989) who uses factor analysis to analyse the "cooccurrence distribution of 67 linguistic features in 481 spoken and written texts of contemporary British English" in two different corpora representing 23 different genres (Biber 1989: 7). The linguistic features represent 16 major grammatical categories, including tense and aspect markers, questions, passives, modals and negation (ibid.). Biber identifies eight text types, or 'dimensions', as he calls them, in his original study (1988), condensing them into five a year later (1989). These dimensions are:

1. Involved versus informational production
2. Narrative versus nonnarrative concerns
3. Explicit versus situation-dependent reference

4. Overt expression of persuasion
5. Abstract versus nonabstract style

To Biber, text types are not singular entities but a continuous scale (Biber 1989: 6). If text types were (mostly) binary to Werlich and Reiss, meaning that texts either exhibit a certain type or not, Biber (1989: 6) finds that "no single dimension is adequate in itself to account for the range of linguistic variation in a language; rather, a multidimensional analysis is required." Biber's text types can—and should—thus, always take all of the five dimensions into account.

As each of the dimensions consists of two feature groups, Biber has analysed the linguistic features of both the 'top' and 'bottom' group. In the case of dimension 1, the top group means 'involved' and the bottom group 'informational', for example. Subsequently, there is a different set of linguistic features for dimensions 1-3 depending on whether the text in question is at the top end of the scale instead of close to the bottom. As such, a text should not include large amounts of frequently appearing features from both ends of the scale at the same time—unless of course it is located near the middle of the scalar dimension. In dimensions 4 and 5, no corresponding bottom group features have, however, been identified. Some of these linguistic features are presented in Table 1 below (for the full list and explanations on the features, see Biber 1989: 8-9). (Biber 1989: 8-9.)

| | 1) Involved vs. informational | 2) Narrative vs. nonnarrative | 3) Explicit vs. situation-dependent | 4) Overt expression of persuasion | 5) Abstract vs. nonabstract |
|---|---|---|---|---|---|
| **Top group** | E.g. private verbs, THAT deletion, contractions, present-tense verbs, 2nd person pronouns, DO as pro-verb, analytic demonstrative pronouns etc. | Past-tense verbs, 3rd person pronouns, perfect-aspect verbs, public verbs, synthetic negation, present-participial clauses | WH relative clauses on object positions, pied-piping relative clauses, WH relative clauses on subject positions, phrasal coordination, nominalizations | Infinitives, prediction modals, suasive verbs, conditional subordination, necessity modals, split auxiliaries, possibility modals | Conjuncts, agentless passives, past-participial clauses, BY passives, past-participial WHIZ deletions, other adverbial subordinators |
| **Bottom group** | Nouns, word length, prepositions, type/token ratio, attributive adjectives, place adverbials | Present-tense verbs, attributive adjectives | Time adverbials, place adverbials, adverbs | - | - |

*Table 1: Biber's five dimensions and their linguistic features (Biber 1989: 8-9)*

To further illustrate how the dimensions work on a scale, presented below is Biber's figure (1989: 12) on the mean scores of dimension 1 for nine different genres. As can be discerned from the figure, face-to-face conversations appear to be the best example of an involved text and official texts that of an informational text. Face-to-face conversations include, for instance, a great deal of present-tense verbs and 2nd person pronouns. Official texts, on the other hand, have a much longer average word length and a higher frequency of nouns. General fiction and prepared speeches, on the other hand, are expected to land somewhere between the two in the middle of the scale. (Biber 1989: 11-12.) The genres presented in Figure 1 could also be placed on the other dimensions. In the case of dimension 3 (Explicit vs. situation-dependent), for example, face-to-face conversations are placed low in the scale due to their situation-dependency, whereas official documents are at the top of the scale as they are much more explicit (Biber 1988: 143).

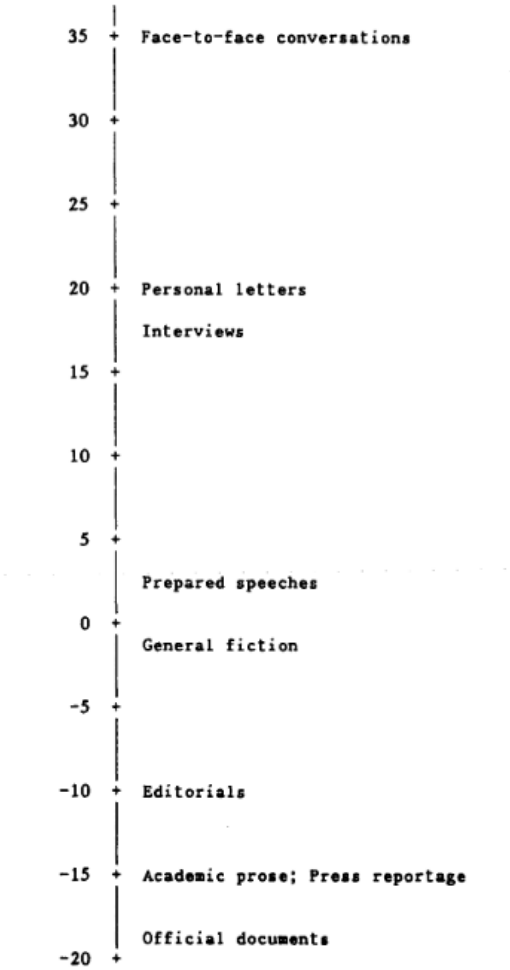



*Figure 1: Mean scores of dimension 1 ('Involved informational production') for nine genres (figure from Biber 1989: 12)*

Corpus-based, quantitative analysis methods have also drawn criticism. Lee (2001: 40-41) draws attention to the fact that corpora do not represent the entirety of a language and that someone has already made a decision on which texts to include in the corpus and which to exclude. Additionally, the corpora have already been categorised using varied external criteria (ibid.). Biber (1993), however, argues in his article on corpus representativeness that corpus construction should be a cyclical process, in which the corpus is repeatedly analysed and studied to identify any sections where more material should be included to make it more balanced. It has, however, been questioned whether a corpus can ever be completely balanced or whether it is even relevant (see Atkins et al. 1992, chapter 7).

### 2.1.3 Discussion

For the purposes of this study, having a set of identifiable linguistic features is a prerequisite. Reiss's functional text types are interesting, but a computable metric is expected to be more objective than a classification based on subjective language functions. Werlich's text types also include linguistic features, but the scalar dimensions used by Biber appear to reflect reality better than purely binary text types. As was mentioned by Reiss, there should always be a dominant text type (Reiss 1977; Fawcett 1997: 104), but most real-life examples are expected to fall into multiple text types anyway, so it makes more sense for the classification to take that into account instead of trying to shoehorn everything into neat little categories. The scalarity also makes it possible to choose as little as three different types of texts to cover the majority of the dimensions, as will be explained in more detail in chapter 3.1.1.

Another benefit of using Biber's scalar dimensions is that he uses arguably more tangible and understandable genres in them. Deciding whether a text is, for example, involved or situation-dependent can be difficult off the top of one's head, but everyone understands immediately what a (prototypical) face-to-face conversation or official document is. There is, arguably, a benefit to being able to present results stating that NMT is better for involved, narrative and situation-dependent texts, and continue that genres X and Y are excellent representatives of these dimensions. To clarify, the purpose of this thesis is to identify which features can be used to predict whether a text will be suitable for NMT but linguistic features are more stable and easier to identify in any given text than contextual features which might change over time. Text types themselves can also, in turn, be used to provide examples on which genres might often include these sorts of linguistic features as, naturally, any non-nonsensical text will always belong to a genre.

## 2.2 Neural Machine Translation

Machine translation (MT) is by no means a new idea. “Mechanical translation” was proposed as early as in the 1940s and the following two decades saw an outright boom of funding for MT research. Breakthroughs in computing power created great expectations and it was largely assumed in the 50s that fully automatic translation would soon be a reality. In the 60s, researchers, however, came to the conclusion that dealing with the ambiguity of real-life language was impossible with the technology and knowledge of the time, and the (in)famous 1966 ALPAC report struck a major nail to the coffin of contemporary MT excitement and funding by stating that “there is no immediate or predictable prospect of useful machine translation” (ALPAC 1966: 32). (Hutchins 2007.)

MT research continued around the world during the following decades, albeit at a slower pace. In the 70s, some projects saw fruition, most notably the Canadian METEO system which translated standardised weather reports between English and French. The systems of the time were mostly rule-based MT (RBMT) engines. RBMT engines are trained manually with detailed information on, for instance, morphological and grammatical rules. By the beginning of the 1990s, corpus-based MT gained prevalence, most notably as statistical MT (SMT). SMT uses large corpora of aligned texts and their translations, focusing initially on individual words. SMT was followed by phrase-based SMT[3] (PBSMT) which learns phrases instead of individual words, which was found to yield better results. (Hutchins 2007.)

PBSMT systems remained state-of-the-art until major breakthroughs were made in neural network training that made neural machine translation (NMT) feasible in the 2010s (Bentivogli et al. 2016: 1). The technology had to first overcome issues concerning computational and resource costs (ibid.) but first demonstrations were soon performed (by e.g. Sutskever et al. 2014; Bahdanau et al. 2014) and, in 2016, NMT was implemented to commercial systems by, for instance, Google (Wu et al. 2016) and Systran (Crego et al. 2016). NMT engines were found to quickly reach the quality of well-established PBSMT engines and even surpass them in some languages (see e.g. the references in this paragraph).

The purpose of this chapter’s introduction to neural machine translation (NMT) is not to provide an exhaustive explanation on how all the different technical or mathematical aspects behind it work (for that, see e.g. Goldberg 2017). Describing a state-of-the-art NMT engine would, in any case, be

[3] PBSMT is used here as an umbrella term for phrase-based, hierarchical and syntactical SMT (similarly to Bentivogli et al. 2016: 1).

somewhat futile as the area is developing in such a rapid speed with new architectures and methods constantly rising and falling (as noted by e.g. Forcada 2017: 8). Therefore, this chapter will focus on giving a basic overview of NMT and its characteristics in sections 2.2.1 and 2.2.2, respectively, and then discuss how those affect this study in section 2.2.3.

### 2.2.1 The basics

Neural machine translation (NMT) is, in a nutshell, corpus-driven MT just like statistical (SMT) and phrase-based MT (PBSMT). The engine is trained using large quantities of bilingual data that include text segments in the source language and their aligned translations. (Forcada 2017: 2.) The training data can amount up to millions of segments (Forcada 2017: 2) or even to trillions, as was the case with Google already in 2007 (Brants et al. 2007: 858). A rule of the thumb with any MT is that the more training data you have, the better the translation (ibid.), and this has been shown to be especially true for NMT (Koehn & Knowles 2017: 1). The thing that separates NMT and the others, however, is what gives NMT its name: *neural networks*.

A neural network is a mathematical model that, in a way, simulates an artificial brain. The network consists of thousands of individual 'neurons' that form connections between each other. When provided with stimuli (i.e. input), the neurons can become either excited or inhibited depending on whether the stimulus is positive or negative. Through their connections, the neurons also excite or inhibit other neurons connected to them, which will either strengthen or weaken the connections. These connections have weights which, in the case of MT, are used to indicate whether a word or a phrase might be translated in a certain way. One single neuron or even a handful are not expected to make any sense, but when grouped up in larger numbers and into multidimensional layers[4] containing hundreds of neural units, they are able to calculate answers to very complicated problems. (Forcada 2017: 2-4.) In addition to language processing, neural networks are currently being used for everything from facial detection (Li et al. 2015) to real-time object recognition (Maturana & Scherer 2015), which are simple tasks to humans but very complex to machines. The key innovation of neural networks is that the network needs no predefined rules on, for instance, which features it should be looking at in the data but learns both the features and their classifiers autonomously from the data (ibid.: 1).

[4] It is these layers working together that make the networks *deep* and give the name to *deep learning* (Forcada 2017: 4).

Here is an extremely simplified example. If a neural network has been given countless examples that sentences beginning with the word “This” tend to be translated into Finnish starting with the word “Tämä”, the network is expected to have very strong connections between the neurons triggered by them. If the training data also happens to include one bilingual segment pair where the first word has been omitted from the Finnish translation for whatever reason (e.g. “This [cucumber is...]” → “Kurkku [on...]”), the network probably includes some connections between the neurons denoting the words “This” and “Kurkku”. They are, however, not nearly as strong as the ones with many repetitions and, thus, the network is able to calculate that the probable translation for “This” is “Tämä”.

The part of the engine that goes through the source text is called the *encoder*, whereas the one that produces the actual translation in the target language is called the *decoder* (Forcada 2017: 7-8). It should be clarified that no NMT encoder will ever actually look at only the first word in either the source or target text as in the previous example. The best encoding architecture has, however, been a matter of some debate. The simplest and first was the sequence to sequence, or seq2seq, architecture (see Sutskever et al. 2014) which was almost immediately extended with an attention mechanism (Bahdanau et al. 2014). In seq2seq, the decoder decodes the individual sentence piece by piece: 1) ‘This’, 2) ‘This cucumber’, 3) ‘This cucumber is’, 4) ‘This cucumber is rotten’. The attention mechanism allows the engine to focus on the segment being translated without forgetting what was said previously like in basic seq2seq (Forcada 2017: 8). Other methods include a convolutional architecture that is able to make generalisations by adding up activations from lower layers (Gehring et al. 2017) as well as a transformer architecture that processes a sequence in parallel using only attention mechanisms (Vaswani et al. 2017). It should, however, be noted that not all architectures have been developed with only translation in mind.

For an individual translator, training and using a personal NMT engine can be somewhat tricky. There are freely available open source toolkits, like OpenNMT[5], but any useful engine requires a great deal of training data and obtaining enough usable and relevant material can be difficult. The actual process of training an engine is also extremely resource-intensive, taking days, weeks or even months depending on the setup, after which even the actual process of translation can be exceedingly slow without specialised hardware. Some of the steps also require in-depth knowledge of computer wizardry (or at

[5] http://opennmt.net/

least the ability to follow detailed guides) that may not be self-evident to most translators. (Forcada 2017: 11-12.)

### 2.2.2 Characteristics of NMT output

NMT has been shown to produce much more fluent text than older SMT models by, for instance, Bentivogli et al. (2016) who looked at word order errors in NMT compared to PBSMT. NMT was found to be much better with verb order (-71% errors) and, to a lesser extent, noun order (-47%). The smallest gains, however, were with prepositions (-18%), negation particles (-17%) and articles (-4%). They discovered that while NMT showed major improvements with word order and, thus, fluency, it was practically just as poor with semantic ordering of adjunct prepositional phrases and the focus of negation, both of which have major implications on the actual semantic meaning of a sentence. (Bentivogli et al. 2016: 8-9.) Similar results were later achieved by Toral & Śanchez-Cartagena (2017).

In Bentivogli et al. (2018), NMT is interestingly shown to make more errors with lexical choice than PBSMT, especially with proper nouns (albeit less overall lexical errors). The authors note that while numerically there were few errors, they were very impactful for the adequacy of the text. In the case of morphology errors, NMT performed significantly better than PBMT, especially with the agreement phenomenon (adjective inflection based on noun) when translating into German. (Bentivogli et al. 2018: 61-64.)

According to Koehn and Knowles (2017: 1), NMT can "completely sacrifice adequacy for the sake of fluency." NMT models require enormous amounts of training data and can produce pure nonsense without enough material but, on the other hand, perform better than previous state-of-the-art engines when provided with high amounts of data (ibid.). NMT engines are generally good at specialising into very specific areas of texts but perform worse than SMT when faced with out-of-domain texts (ibid.: 2). The engine is, thus, good in situations that emulate the conditions under which it has been trained but, subsequently, has difficulties when those parameters are changed significantly (ibid.: 10). A solution to this might be to deliberately build the engine from as varied a set of material as possible (ibid.), but that might lead to it being less useful in more specific areas. This has been answered by first training an engine using more general data and only then fine-tuning, or *adapting*, it with domain-specific material, which has been shown to increase translation quality (see e.g. Luong & Manning 2015).

There are some differing opinions on whether NMT is good with rare words or not, which, in a way, is another out-of-domain situation. It should be noted that previous MT models used large external

vocabularies, whereas NMT generally only consists of the training corpus (Koehn & Knowles: 6). Multiple studies (e.g. Sutskever et al. 2014; Bahdanau et al. 2014; Wu et al. 2016; Bentivogli et al. 2018) refer to NMT having difficulties with rare words in comparison to PBSMT. Some ways to address this have been proposed, most notably the byte-pair encoding system (see Sennrich et al. 2016) that allows the decoder to identify individual sub-word character sequences, or morphemes, of a word (Koehn & Knowles 2017; for other methods, see e.g. Arthur et al. 2016 and Luong et al. 2014). Koehn and Knowles (2017: 1) note that NMT engines operating on a sub-word level are, at least in their study that uses a byte-pair encoding model, in fact better than SMT with infrequent words, although the margin grows much smaller when dealing with highly inflected word classes such as verbs. The issue of fluency over adequacy is, however, also apparent with rare words, and an example is given where, in the sentence "*[The] police closed in on him",* the engine chose to translate the German verb *einkesselte* (closed in on) as *stabbed* (Koehn & Knowles 2017: 6), which arguably sounds very fluent but makes the sentence slightly more macabre than intended. For a more detailed description on why NMT does these types of mistakes, see the introduction to Arthur et al. (2016).

NMT has also traditionally performed worse with sentences that are longer than around 60 words (Koehn & Knowles 2017), although some studies have not been able to repeat this finding (see e.g. Bentivogli et al. 2018: 58). This dispute can be partly explained with the use of different encoder-decoder models, as for instance attention models were found to greatly increase an engine's capability of translating longer sentences (Koehn & Knowles 2017: 6). The study by Koehn and Knowles (2017) does use an attention mechanism, but the one by Bentivogli et al. (2018) also refers to using a bidirectional encoder (i.e. one that codes the sentence in both directions) that might explain the better performance. It should be noted that there is still great variance between different NMT systems (Toral & Śanchez-Cartagena 2017: 9) which might explain the differences between studies.

For a translator or post-editor, the output of an NMT engine is in many ways different from an SMT engine. NMT was found to be much more fluent than previous MT iterations (Koehn & Knowles 2017: 1; Forcada 2017: 13; Toral & Śanchez-Cartagena 2017), but it often makes semantic errors and can, for instance, translate a country's name (like *France*) as a different one (*Germany*) (a practical example noted separately by both Forcada 2017: 11 and Arthur et al. 2016: 1). As was noted previously, these kinds of low-frequency but semantically important words are a noted challenge for NMT and some improvements have been proposed. For a translator, spotting these kinds of errors can be difficult in an otherwise fluent text but vital for preserving the meaning of the source text (Forcada 2017: 11-12).

To summarise, NMT output is much more fluent than the one produced by previous models but faces issues with semantics. These errors might be a very small part of a text but sometimes turn its meaning on its head and spotting them can be tricky if a text appears fluent and otherwise correct. Another potential issue were too long sentences and rare words, although there was some dispute as to whether this is true or not. NMT was also found to face issues with lexical choice, especially concerning proper nouns. The engines themselves perform best in situations that are similar to those under which they have been trained but face difficulties when presented with out-of-domain material.

### 2.2.3 Discussion

For the purposes of this study, it was important to choose an NMT engine that was able to adequately translate texts in various text types. As such, the engine had to be a generalist to avoid out-of-domain situations where NMT engines were noted to face difficulties. Building a new engine was not a possibility due to the amount of training data required for each of the different genres and time constraints of the thesis. As such, two existing engines were chosen for the study: Google's Cloud Translate and Lingsoft's NMT engine, which are presented in slightly more detail in chapter 3.3 with an explanation on why the two were specifically chosen for the study. Using two separate engines should help alleviate the issue of a single engine distorting the results during a time that the setups of NMT engines are still in flux. A more specialised MT engine would obviously be expected to perform better than a general one, but in this case, the two should be enough to give preliminary results on which texts and linguistic features appear to work best with NMT.

It should also be noted that as the engines and their compositions are still changing rapidly, there are certain features I will identify in this study that may or may not actually be accurate for all different NMT engines running varied compositions, or if and when new breakthroughs are made in the future. The issue with sentence length, for example, was noted to diminish with newer architectures. As such, the results might not be as generalisable as one would hope but should still provide some indicative results. The evaluation method used in this study is also something that can be used in future studies.

Regarding NMT's potential issues and characteristics, they can be broadly distributed into two separate categories: features of the source text (domain-specificity, sentence length, rare words) and features in the translation (fluency, adequacy, lexical choice, semantic errors). Measuring whether these are issues of NMT specifically was somewhat difficult in this study, as the same data sets would had to have been translated using PBSMT or SMT as well and see whether the results where different or similar compared

to NMT. Some of the error types are, however, included in the error typology (see chapter 3.2.1): domain-specificity is included in term errors, fluency in grammar and idiomaticity, and lexical and semantic errors in mistranslations. Sentence length will be analysed separately in chapter 4.3. Adequacy was excluded from the typology as this study did not want to evaluate the *quality* of the translations, which is what adequacy ultimately is in relation to whatever the text will be used for. Studying rare words would also have been difficult, as there is no way to know the exact composition of the material on which commercial engines are trained, meaning that it is impossible to know which words are actually rare to the engine and which are not.

# 3 Material and method

This chapter focuses on the material and methods used in the study. First, chapter 3.1 discusses material selection. Chapter 3.2 introduces the error typology used in the study and briefly discusses automatic evaluation and the metric chosen for this study. Finally, chapter 3.3 will discuss the two NMT engines used in the study.

## 3.1 Material

In this chapter, it is first discussed in chapter 3.1.1 how Biber's five dimensions can be covered with the least number of genres. The material chosen for the study is then presented in chapter 3.1.2.

### 3.1.1 Covering the five dimensions

As was discussed in chapter 2.1.2, Biber's (1988; 1989) text type categorisation consists of five dimensions:

1. Involved versus informational production
2. Narrative versus nonnarrative concerns
3. Explicit versus situation-dependent reference
4. Overt expression of persuasion
5. Abstract versus nonabstract style

In his original study, Biber (1988) used 23 different subgenres, as he calls them, with a combined total of 481 texts and showed how each of them is positioned on every dimension (see Appendix 1). Telephone conversations, for instance, received a score of 35[6] on dimension 1, -2 on dimension 2, -5 on dimension 3 and so forth. It should be noted that the dimensions do not use standardised scales and, for instance, the highest score a genre receives in dimension 1 is 35, whereas in dimension 2 only 7 and in dimension 4 a meagre 3. As such, the genres presented in both ends of dimension 1 are expected to be further apart from each other than in dimension 4, but only regarding the relevant linguistic features of each dimension that were presented in Table 1 in chapter 2.1.2. In other words, each of the dimensions uses different criteria (i.e. linguistic features) to place the subgenres on the scale. (Biber 1988: 67, 128-155.)

[6] The scores have been rounded here to the nearest full integer for simplicity. For the accurate values, see Biber (1988: 128-155).

As each genre can be placed in each of the five dimensions, it is possible to cover most of them choosing only three of Biber's genres: **fiction**[7], **professional letters** and **official documents**. As the chosen material was, however, not measured to objectively match the linguistic criteria set by Biber (1988), it was deemed misleading to expect the material to fully match the one Biber used in his study. As such, a slight simplification was made by using the top and bottom group separation presented in chapter 2.1.2 and by adding a new middle group to represent values close to the middle of the scale. It was then possible to place the three abovementioned genres to the dimensions, as presented in Table 2 below. The original values of each genre are presented in brackets.

| | Dimension 1 | Dimension 2 | Dimension 3 | Dimension 4 | Dimension 5 |
|---|---|---|---|---|---|
| **Top group** | - | Fiction (6) | Official documents (8.4), professional letters (7.7) | Professional letters (3.4) | Official documents (4.8) |
| **Middle group** | Fiction (-2), professional letters (-3) | - | Fiction (-3.3) | Fiction (1), official documents (0) | Professional letters (0.2) |
| **Bottom group** | Official documents (-19) | Professional letters (-2.3), official documents (-2.8) | - | - | Fiction (-3) |

*Table 2: Simplified placing of three subgenres in each of the five dimensions and their scores*

The top group of dimension 1 and bottom groups of dimensions 3 and 4 are underrepresented, but relevant material (namely broadcasts and telephone conversations) and their translations needed to fill the gaps was deemed too difficult to acquire to fit the time constraints of the thesis. As can be discerned from the scores, the scores for texts in the top group tend to be higher than their opposites at the other end of the scale, with the exception of dimension 1. As such, making a difference between the middle and bottom group was sometimes difficult, as for instance fiction could be placed in either group in dimension 3. Fiction was ultimately placed in the middle group as the original scale continued with telephone conversations and broadcasts with scores of -5.4 and -9, respectively.

[7] It should be noted that while Biber differentiates between different types of fiction (general, mystery, science, adventure and romantic), the differences on each dimension are minimal and were, thus combined into a single category.

### 3.1.2 Material used in the study

Three separate corpora were collected for the analysis. For the **Professional Letters** (PL) corpus, a total of 27 exchanges of letters were chosen from EUR-Lex. All the chosen letters included a Finnish translation, but the documents do not specify whether English is always used as a source language. As the material is picked from a single source, it is, however, expected to follow a similar structure regardless of the language they have originally been written in. It should be noted that each individual 'professional letter' was always an exchange of letters and generally included more than one consecutive letter on the same topic. Presented below in Example 1 is the beginning of a letter and its reference translation that were used in the material.

| English | Finnish reference translation |
|---|---|
| Sir,<br><br>I have the honour to refer to your attached letter ref D/001782 of 17 September 2003 on the establishment of official relations between our two organisations.<br><br>I agree with your proposal to reinforce relations between the Office international des épizooties and the Commission of the European Communities based on the following: | Arvoisa komissaari<br><br>Haluan viitata 17. syyskuuta 2003 päivättyyn kirjeeseenne (viite D/001782), joka koskee virallisten suhteiden luomista järjestöjemme välille.<br><br>Hyväksyn ehdotuksenne Maailman eläintautijärjestön ja Euroopan yhteisöjen komission välisten suhteiden lujittamisesta seuraavien periaatteiden pohjalta: |

*Example 1: Professional letter in the material*

For the **Official Documents** (OD) corpus, a total of 20 texts were chosen from EUR-Lex, once again with Finnish translations included. The texts are all preparatory documents (sector 5 on EUR-Lex) and consist of ten resolutions and ten reports. Compared to the material chosen for professional letters, the official documents were much longer in length and there was thus no need to include as many of them. Two separate categories were chosen to provide a slightly better picture of the broad genre. Presented below in Example 2 are two examples of the beginnings of a proposal for regulation and a report.

The **Fiction** (Fic) corpus is from OPUS, an open source parallel corpus (see Tiedemann 2012). The material consists of Sir Arthur Conan Doyle's The Hound of Baskerville and its Finnish translation by Yrjö Weilin from 1904. The OPUS corpus also included Lewis Carroll's Alice's Adventures in Wonderland, but it was deemed to be a bad fit for a MT study due to its frequent use of poems and non-standardised vocabulary. Presented below in Example 3 is the beginning of The Hound of Baskervilles.

| English | Finnish |
|---|---|
| Resolution (EU) 2016/2155 of the European Parliament<br><br>of 27 October 2016<br><br>with observations forming an integral part of the decision on discharge in respect of the implementation of the budget for the ENIAC Joint Undertaking for the financial year 2014<br><br>THE EUROPEAN PARLIAMENT,<br><br>having regard to its decision on discharge in respect of the implementation of the budget of the ENIAC Joint Undertaking for the financial year 2014,<br><br>having regard to Rule 94 of and Annex V to its Rules of Procedure,<br><br>having regard to the second report of the Committee on Budgetary Control (A8-0264/2016), | Euroopan parlamentin päätöslauselma (EU) 2016/2155,<br><br>annettu 27 päivänä lokakuuta 2016,<br><br>joka sisältää huomautukset, jotka ovat erottamaton osa päätöstä vastuuvapauden myöntämisestä ENIAC-yhteisyrityksen talousarvion toteuttamisesta varainhoitovuonna 2014<br><br>EUROOPAN PARLAMENTTI, joka<br><br>ottaa huomioon päätöksensä vastuuvapauden myöntämisestä ENIAC-yhteisyrityksen talousarvion toteuttamisesta varainhoitovuonna 2014,<br><br>ottaa huomioon työjärjestyksen 94 artiklan ja liitteen V,<br><br>ottaa huomioon talousarvion valvontavaliokunnan toisen mietinnön (A8-0264/2016), |

*Example 2: Official document in the material*

| English | Finnish |
|---|---|
| Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table.<br><br>I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before.<br><br>It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." | Herra Sherlock Holmes, joka tavallisesti nousi hyvin myöhään ylös aamuisin, paitsi niissä kylläkin useissa tapauksissa, jolloin hän oli valvonut koko yön, istui aamiaisella.<br><br>Minä seisoin matolla tulisijan edessä pitäen kädessäni keppiä, jonka eräs edellisenä iltana luonamme käynyt herra oli unohtanut.<br><br>Se oli jokseenkin soma ja tukeva, se oli varustettu sipulinmuotoisella kädensijalla ja näytti oikealta "tuomarin sauvalta." |

*Example 3: Fiction in the material*

## 3.2 Methods

Next, the error typology used in the qualitative part of the study is introduced in chapter 3.2.1. It is followed by a brief introduction to automatic evaluation and the metric chosen for the study, LeBLEU, in chapter 3.2.2.

### 3.2.1 Error typology for sentence-level analysis

As the second part of the study is a qualitative sentence-level analysis, an error classification was needed. For that, I used the DQF-MQM error typology (TAUS 2019) as a basis. The DQF-MQM typology consists of eight error types with several subcategories (see ibid.):

1. Accuracy
2. Fluency
3. Terminology
4. Style
5. Design
6. Locale convention
7. Verity
8. Other

Some of these can be easily discarded. Design includes sentence length as a subcategory, which is already analysed separately, and all its peers (local formatting, markup, truncation/text expansion) are irrelevant for the study. Locale conventions and verity can also be discarded, as a MT engine is not expected to use external knowledge to adapt a text to match local conventions or expectations[8]. The category other can also be discarded, as the remaining categories were deemed sufficient for the error categorisation. This leaves out four categories: accuracy, fluency, terminology and style. They, however, also include irrelevant subcategories. There is no need to be looking for, for instance, improper TM matches as no TM was used. It is also impossible to evaluate the consistency of the texts as the sentence-level analysis did not include them in their entirety.

The final error categorisation is, thus, the following:

1. Accuracy
    1.1. Addition
    1.2. Omission
    1.3. Mistranslation
    1.4. Untranslated
2. Fluency
    2.1. Grammar
    2.2. Spelling
    2.3. Punctuation

[8] Unless of course it has been trained using material that, for instance, always changes dates from mm/dd/yyyy to dd.mm.yyyy.

3. Terminology
4. Style
    4.1. Awkward
    4.2. Unidiomatic

A separation between additions, omissions and mistranslations was deemed suitable, as was including separate categories for term errors and mistranslations. The latter distinction was, however, discovered to be somewhat problematic during the analysis, as will be discussed in chapter 4. There were only a few untranslated or misspelled words and punctuation errors, but including the three categories was still deemed justified. Making a separation on the structural level between grammar (incorrect), unidiomaticity (grammatically correct but makes little sense) and awkwardness (correct and makes sense but with unnecessarily difficult sentence structures) was also deemed successful.

### 3.2.2 Automatic evaluation

The concept of automatic evaluation revolves around the idea that "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al. 2002: 1). In practice, this means comparing a machine translated text to a translation of the same text by a human translator. An evaluation metric is used to calculate and give a score to the MT output using differently weighted criteria. This enables a researcher or MT developer to quickly gain indicative results about whether a certain adjustment improved the MT quality or just to compare two or more MT engines (Papineni et al. 2002) in a way that is faster, easier and cheaper than by using human evaluators (Banerjee & Lavie 2005: 1). The tools themselves are, however, calibrated using human evaluators (Babych 2014: 465). Babych (2014: 468-469) discusses a need to recalibrate the metrics whenever something as simple as the target language changes, with MT type as an extreme example. In other words, if a metric has been calibrated using RBMT, it will practically always rate texts translated with such engine the best. For a more in-depth analysis on what exactly is automatic evaluation, see for example Papineni et al. (2002) on the origins of BLEU and Banerjee and Lavie (2005) on METEOR.

**BLEU**, or the *Bi-Lingual Evaluation Understudy* (Papineni et al. 2002: 1), is by far the most popular automatic evaluation metric (Reiter 2018: 394; Shterionov et al. 2018: 5). When Google, for instance, introduced its new NMT engine in 2016, BLEU was the sole evaluation metric used in the study (Wu et al. 2016). BLEU looks at the n-grams of the candidate translation and compares them to the reference translation, counting the matches independently of their position (Papineni et al. 2002: 2). It focuses on translation length, which should not be too long or too short, and the n-grams of the MT output and the

reference translation (Shterionov et al. 2008: 5-6). The metric then scores the translation with a weighted average that ranges from 0 to 1, where 1 means complete equivalence with the reference translation and is, as such, only a hypothetical gold-standard (Papineni et al. 2002: 5). The original study appears to show a very high correlation between reference human evaluators and the BLEU scores (Papineni et al. 2002: 7-8). The reason why the metric remains so popular after 15 years, is that it works with any language and is fast to use (Shterionov et al. 2018: 5). BLEU has also worked as a basis for other derivative metrics, such as NIST (see Doddington 2002), which finetuned the BLEU metric with a few additional metrics. Some studies use several metrics (e.g. Rautio & Koponen 2013), but BLEU appears to be the most widely used.

BLEU has, however, been criticised since its inception. A clearly identified fault in BLEU is that it is unable to reliably score individual sentences, focusing instead on the entire text (Lavie & Denkowski 2009: 1). According to a heavily-referenced article by Callison-Burch et al. (2006: 1), the MT community is "overtly reliant" on BLEU and that an increase in BLEU score might not translate to an actual improvement in translation quality. A more recent structured review by Reiter (2018) came to the conclusion that "the evidence supports using BLEU for diagnostic evaluation of MT systems (which is what it was originally proposed for), but does not support using BLEU outside of MT, for evaluation of individual texts, or for scientific hypothesis testing" (ibid.: 393). He criticises the metric for the lack of testing with real-world applications and its technological "biases" that might affect how it scores NMT systems (ibid.: 399-400).

As this study focuses specifically on NMT, its characteristics needed to be taken into account in the metric selection. There are studies (e.g. Shterionov et al. 2018) showing that older metrics, such as BLEU and TER, tend to underestimate the quality of NMT. According to Shterionov et al. (2018: 6), NMT produces much more variation in terms of sentence length and word choices, which are things that traditional evaluation metrics tend to base their scores on, naturally with some variation. As an NMT-based system is able to look past individual sentences and their n-grams, in a way looking at the bigger picture like a human translator, it is more likely to make more unexpected choices that might fool the evaluation metric to think that something is incorrect. NMT systems have also been found to be more creative in their use of synonymous phrases. This might result in the evaluation metric thinking that the NMT has used an incorrect translation, even if the semantic meaning is the same. (Shterionov et al. 6-7.)

The study also found that human evaluators always gave better scores to translations by NMT engines than those created by a PBSMT engine, even when BLEU scored them similarly. The time required to post-edit NMT was, with the exception of the English-Chinese (Simplified) language pair, also found to

be shorter for PBSMT. The study also concluded that BLEU underestimated 47% of NMT output, while only 17% of PBSMT output, which can be seen as a proof of the biases discussed by Reiter (2018: 400). (Shterionov et al. 2018: 10-15.) The correlation of BLEU scores is also known to drop in higher quality texts (Babych 2014: 469), which can be a major fault when scoring state-of-the-art MT output.

**LeBLEU**, meaning *Letter-Edit-BLEU* or *Levenshtein-BLEU,* by Virpioja and Grönroos (2015) takes the original BLEU and adds fuzzy n-gram matching. Instead of seeing n-gram correctness as binary: either correct (1) or incorrect (0), LeBLEU turns it into a scale. This allows the metric to make the distinction that is obvious to humans that, for instance, the phrases “black-and-white” and “black and white” are more similar to each other than “black-and-white” and “cucumber”. So instead of marking the unhyphenated one as incorrect, the metric is able to presume that this is most likely the same n-gram after all and provide it with a score that is probably just a bit less than a full match. The study also shows that, unlike BLEU, LeBLEU is highly correlative with human judgement already on a sentence-level, which is important for the second part of the analysis. (Virpioja & Grönroos 2015: 411-412.)

Fuzzy n-gram matching appeared to be a simple but major improvement for evaluating not only NMT but also morphologically complex languages, like Finnish, which was chosen for the study. Considering the criticism, it was hard to justify using BLEU in a study focusing specifically on NMT, even while taking into account its prevalence in the field. LeBLEU was clearly a more promising option.

## 3.3 NMT engines used in the study

The material described in chapter 3.1.2 was translated using two separate NMT engines. The first was Google’s commercial Cloud Translation (not to be confused with the free Google Translate) and the second an NMT engine by Lingsoft, a Finnish language technology and management company. Due to their proprietary nature, there is little documentation available on the specific setup of the engines or statistics on the material they have been trained on. There is some information available on what Google’s engine looked like in 2016 (Wu et al. 2016), but undoubtedly it had changed by the time the material in this study was entered into it in early 2019. Lingsoft’s engine is based on the open source Marian engine[9], uses sub-word unit vocabulary such as byte-pair encoding (to reduce vocabulary size and handle inflection morphology) and has been trained on a dataset that includes material from their customers, namely public administration, electronics, automotive, software and medical texts.

[9] https://marian-nmt.github.io/

Google's Cloud Translation was chosen due to Google's apparent interest in developing and implementing state of the art MT features (Wu et al. 2016) and its massive data pool on which it has trained its engine (Brants et al. 2007: 858). Google Translate is obviously also a brand and feature known by a great deal of people and it was interesting to see how competitive the commercial version was. Lingsoft's engine was included in the study for two reasons: 1) as NMT systems are still somewhat varied in their features, having two systems was expected to make it easier to draw conclusions that might be more relevant to NMTs in general instead of just to a single engine. 2) Finnish is arguably a very small language and it is not expected to be the focus of much research or effort in mainstream NMT studies. As such, an engine that has been developed in Finland over a long period of time was expected to take the peculiarities of Finnish better into account and it was interesting to see whether this would reflect in the translation quality.

# 4 Analysis

This chapter presents the results of the analysis. First, in chapter 4.1, corpus level scores are introduced and their implications to Biber's text type dimensions are discussed. Chapter 4.2 focuses on segment-level scores and discusses which error types were most prevalent for each genre and engine. Finally, in chapter 4.3, the relation between average sentence length and evaluation scores is briefly discussed.

## 4.1 Corpus level analysis

This chapter discusses the corpus level LeBLEU scores in chapter 4.1.1 and how they fit Biber's dimensions in chapter 4.1.2.

### 4.1.1 LeBLEU scores

Presented below in Table 3 are the corpus level LeBLEU scores for the six data sets.

| Corpus | NMT engine | LeBLEU score |
|---|---|---|
| **Fiction (Fic)** | Google | 0.507769 |
| | Lingsoft | 0.452538 |
| **Professional Letters (PL)** | Google | 0.688428 |
| | Lingsoft | 0.661735 |
| **Official Documents (OD)** | Google | 0.683341 |
| | Lingsoft | 0.648748[10] |

*Table 3: Corpus level LeBLEU scores*

Looking at the scores received by individual text types, it is obvious that Fiction (Fic) was by far the most difficult genre for both engines. While Google performed better with it than Lingsoft, the results are still relatively weak compared to the Professional Letters (PL) and Official Documents (OD) corpora.

The results are also interesting as Lingsoft was expected to perform better with the peculiarities of Finnish, as was discussed in chapter 3.3, but Google was still better with the three genres. A possible answer to this could be that Google just has much more material to train their engines with. As was also discussed in chapter 3.3, one of the key innovations with neural networks was that it is able to work autonomously with the data to learn which features and classifiers it should be looking at (Maturana & Scherer 2015: 1), which could mean that language specific knowledge and adaption might no longer be as important as with PBSMT. As such, it might really be that training data is king with NMT. Regardless,

[10] It should be noted that, after the analysis, some of the material was discovered to have been missing from Lingsoft's OD corpus and had thus not been translated. This might make the results between the two engines slightly less comparable regarding the OD corpus, but the material was not rerun due to time constraints and as the result achieved by Lingsoft's NMT engine appeared to still match its expected value relatively well.

in chapter 4.2.3, some of the differences between the two engines that were discernible in the analysis and their most common error types are discussed.

The reason why fiction appears to perform relatively poorly could be the subject of another thesis entirely, but there might be some relatively simple answers. First, as was discussed in chapter 3.2.2, automatic evaluation is generally not interested in whether a translation is semantically equivalent or at least close to the source text, but whether it is similar to the *reference translation*. If there is, thus, greater variance between potential translations of a source text, as there is in translations of fiction, it logically follows that those texts are expected to perform worse with automatic evaluation than those that generally always follow similar patterns and sentence structures, as is the case with the two EU corpora. Having analysed some of the segments of the Fic corpus more closely, it was obvious that the original translator had only rarely chosen to do a direct or literal translation of the source segment, opting instead to often shift something between nearby sentences, move the focus of the segment for more clarity or even add something new for dramatic or whatever purposes, as can be seen in Example 4 below.

| Source text | NMT output | Original translation |
|---|---|---|
| That is his mark. | Tämä on hänen merkkiään. | Siihen he ovat jättäneet merkkinsä. |

*Example 4: Creativity in the Fic corpus*

Here, the human translator has opted to translate “That is his mark” as “They have left their mark there”, whereas the NMT engine has produced a literal translation as expected. This leads to the second point: NMT, and MT in general, appear to prefer literal translations. It does not care whether the text appears dramatic or understandable to the human reader but aims to emulate the material it has been trained with to produce a translation. Below is Example 5 from the PL corpus to showcase this. The reference translation emulates the grammatical relations between the actors in the sentence perfectly.

| Source text | NMT output | Original translation |
|---|---|---|
| For the Government of the Republic of Indonesia this legal framework is based on Indonesian Law No 2 of 1982 dated 25 January 1982 concerning the Ratification of Convention on Special Mission, 1969. | Indonesian tasavallan hallituksen osalta tämä oikeudellinen kehys perustuu Indonesian lakiin nro 2, joka on annettu vuonna 1982, erityisedistystä koskevan yleissopimuksen ratifioinnista 1969. | Indonesian tasavallan hallituksen osalta nämä oikeussäännöt perustuvat 25.1.1982 annettuun Indonesian lakiin N:o 2, joka koskee erityisoperaatioita koskevan vuoden 1969 yleissopimuksen ratifiointia. |

*Example 5: Segment in the PL corpus*

Nothing new is added and nothing is left out, although the translation the date is moved slightly to make the sentence more fluent. The NMT output includes a few term errors but the engine has again done a literal translation and manages to keep all grammatical relations intact.

These two points make it clear that the automatic methods generally used to evaluate MT output greatly prefer literal translations and, as such, are biased against translations of fictional text. The segment level analysis that will be discussed in chapter 4.2 should circumvent the issue by focusing on actual errors in the translations and disregarding things such as Example 4 of the Fic corpus, where the engine cannot be criticised for not including additional embellishments by the human translator. There is naturally also the question of whether MT is useful for the translation of fictional texts if it only opts to do literal translation, but that is outside the scope of this thesis. On the other hand, it would be interesting to see how an NMT engine that has been trained solely on fictional texts and their translations would perform and whether it would in fact opt to do more creative translations than the current engines that are generally trained using material that is similar to the EU corpora.

Regarding the PL and OD corpora, both engines were able to reach much better results. Google was slightly better with both, with a greater margin with the OD corpus (although this might be due to the partly missing material). Interestingly, PL was the best corpus for both engines, although not by much. It is also interesting that both engines received scores that were relatively close to each other with the PL and OD corpora, especially compared to the Fic corpus. This might be due to the fact that both corpora originated from EUR-Lex, which has been widely used for MT training as all of its material is freely available for use. This might have, arguably, made it slightly less ideal for this study and it would be interesting to see whether a similar corpus from another source would yield similar results.

### 4.1.2 Back to the dimensions

So how do the scores look like when placed into Biber's dimensions? As was discussed in chapter 3.1.1, the point of choosing the three genres used in this study was to see which of the dimensions appear to correlate the best with their scores. If they did, these dimensions could then be used to make generalisations that could, in turn, be used to predict the machine-translatability of other genres. Presented below is Table 4 where the relative LeBLEU scores have been placed according to the positions of their representative genres in Table 2 in chapter 3.1.1. For this table, a simple average between the two engines was used (Fic: 0.4801535; PL: 0.675082; OD: 0.666045), which was then rounded to the third decimal for better readability.

| | Dimension 1: Involved vs. Informational | Dimension 2: Narrative vs. Non-narrative | Dimension 3: Explicit vs. Situation-dependent | Dimension 4: Overt persuasion | Dimension 5: Abstract vs. Non-abstract |
|---|---|---|---|---|---|
| Top group | - | 0.480 | 0.666, 0.675 | 0.675 | 0.666 |
| Middle group | 0.480, 0.675 | - | 0.480 | 0.480, 0.666 | 0.675 |
| Bottom group | 0.666 | 0.675, 0.666 | - | - | 0.480 |

*Table 4: Evaluation scores of the three genres in Biber's dimensions*

Looking at the five dimensions, some clearly show stronger correlation than others, although more genres would be required to draw ultimate conclusions. But looking at dimensions using these three genres, dimensions 2 and 3 appear to show relatively strong correlation: In dimension 2, narrative texts received much worse results than non-narrative, whereas in dimension 3, explicit texts appeared to translate better than those in the middle group. Dimensions 1 and 4 are slightly more ambiguous and would require more genres to see how their top and bottom groups, respectively, would perform. In dimension 5, the top (abstract) and middle groups appear to have performed better than the bottom group (non-abstract), but the correlation is not as strong as in dimension 2 and 4.

To summarise, based on the three genres studied in this thesis, the results appear to show that non-narrative texts are more suitable for NMT than narrative texts and that explicit texts are more suitable than situation-dependent. More genres should, however, be studied to be able to draw more wide-reaching conclusions. In the next chapter, I will look at the individual genres and see which error types were most prevalent to each.

## 4.2 Segment level analysis

For this section of the analysis, the individual segments were analysed with the Appraise MT error classification package (see Federmann 2018) using the error categorisation described in chapter 3.2.1. As was mentioned in chapter 3.2.2, LeBLEU scored both the system (presented in Table 3) and its individual segments. These scores varied between anything from 0.000000 (LeBLEU found nothing similar between the two) to 1.000000 (identicality). It should be noted that not all the scores were necessarily valid. For instance, in situations where the NMT had translated "ANNEX" as "Liite" instead of "LIITE" in capitals, it was always given a score of 0.000000. In retrospect, this could have been avoided by evaluating the material case insensitively. Not all 0.000000 scores were false, however, as can be seen in Example 6 below. Here, the NMT had translated "Yours truly" literally as "Truly/really" instead of using a more idiomatic and established alternative: "Respectfully". This example would have been

classified as a mistranslation. In situations where the engine had translated a word using a viable synonym, such as “Ystävällisesti” (“Kindly”), no errors were, however, marked.

| Source text | NMT output | Original translation |
|---|---|---|
| Yours truly, | Todella | Kunnioittavasti, |

*Example 6: Valid 0.000000 score*

It should also be noted that not even good scores always told the whole story. As was described in chapter 2.2, NMT sometimes faces issues where the semantic meaning of a sentence might be lost over an error that is minute to automatic evaluation. Presented below in Example 7 is one such case in the PL corpus. Here, the “exchange of letters […] on cereals” has been translated as “exchange of cereals” but the translation still receives a score of 0.767816 which is much higher than the system level score that the PL corpus received. This example would have been classified as an omission and a mistranslation.

| Source text | NMT output | Original translation |
|---|---|---|
| **EXCHANGE OF LETTERS** between the European Community and the Republic of Argentina **on cereals** | Euroopan yhteisön ja Argentiinan tasavallan **välinen viljanvaihto** | Euroopan yhteisön ja Argentiinan tasavallan **viljaa koskeva kirjeenvaihto** |

*Example 7: Semantic error in segment with high score*

A matter that had to be decided was choosing a selection of segments that provided interesting results when analysed and, above all, variation between the genres. A brief pilot was run on Lingsoft’s PL corpus to see what the ideal score would be where the NMT still had some issues but was able to translate legible sentences. Choosing the absolute worst segments was deemed counterproductive, as there must be something similar between the source and the translation to be able to classify where and how things go wrong. Next, an attempt was made to categorise segments with a minimum score of 0.2, but the issue remained the same. The final attempt was to categorise segments that had a score equal to the system score: in the case of Lingsoft’s PL corpus, 0.661735. This yielded much better results. Mistranslations were still common, as was expected, but the spectrum of different error types became much larger. Looking at the errors in translation with an average score was also expected to yield somewhat generalisable results about each genre. As such, 30 segments with scores starting from the system average were picked from each of the three sets for the two engines for a total of 180 segments. For Google’s PL corpus with the system score of 0.688428, the chosen 30 segments had scores between 0.688664 and 0.703129, for example.

The findings from the error categorisation analysis are discussed in the following two chapters. Chapter 4.2.1 focuses on the three different genres and their differences. Finally, chapter 4.2.2 looks at the differences between the two engines.

### 4.2.1 Frequent errors per corpus

Presented on the next page are three graphs, one for each of the three genres, that show how the error types were spread out between the genres. Looking at the results, there are some similarities between, for instance, the OD and PL corpora, which were expected to be closer to each other than the Fic corpus, but each has their own discernible features. It should be noted that the error types were categorised as binary options in each segment: either an error type was present (1) or it was not (0). As such, a segment might have included multiple major mistranslations, but it was still categorised similarly to a segment that only included one minor mistranslation. Thus, the statistics portray the frequency of the error types **per segment**, not which ones were numerically the most frequent.

Mistranslations were by far the most prevalent error type, and in retrospect the category should have been spread into a few separate error types. There is arguably a major difference between, for instance, semantic errors where the engine has chosen a translation for a word that is wrong in that particular context (see Example 8) and cases where the engine has misunderstood the grammatical relations between different actors in a sentence or has failed to portray them accurately in the translation.

| Source text | NMT output | Original translation |
|---|---|---|
| It is my **business**, and not yours. | Se on minun **yritykseni**, ei sinun. | Se on minun **asiani**, eikä teidän. |

*Example 8: Semantic error*

It was also sometimes difficult to discern between term errors and mistranslations. In the case of the Fic corpus, no term errors were identified in the analysed section of the material, whereas in the PL and OD corpora, term errors amounted to 14% and 18% of the all errors, respectively. Without venturing too deeply into the conundrum of determining what is a term and what a *regular* word, in this study, a term was simply interpreted as a "domain-specific word" as per to the original DQF-MQM typology by TAUS (2019). In practice, this meant that the Fic corpus, for obvious reasons, had very few instances that were considered to be domain-specific terms, although there were some grey areas. Can it, for instance, be considered a term error, if "intolerant (eyes)" is translated as a medical condition or "clutch (in which it held us)" as the clutch in a car? Regarding these two examples, both were simply marked as mistranslations, although one might argue that, even though the words were not specific to the domain

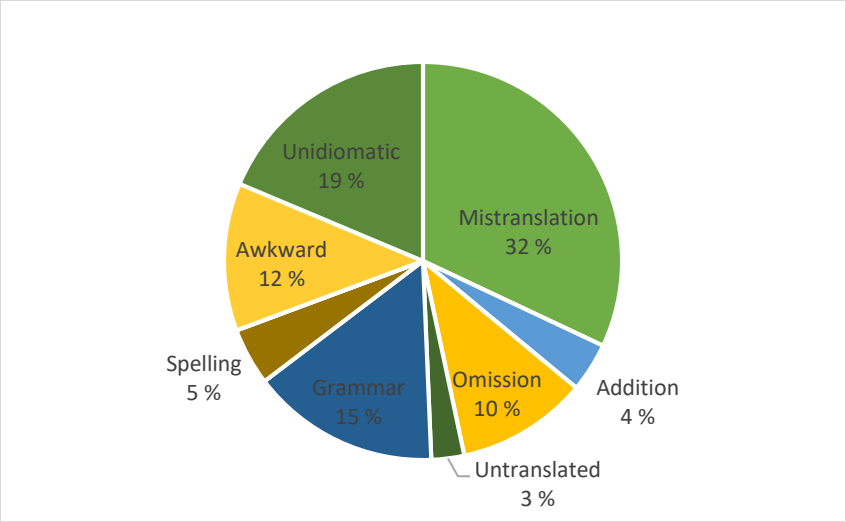



*Figure 2: Error type frequencies for the Fic corpus*


*Figure 3: Error type frequencies for the PL corpus*

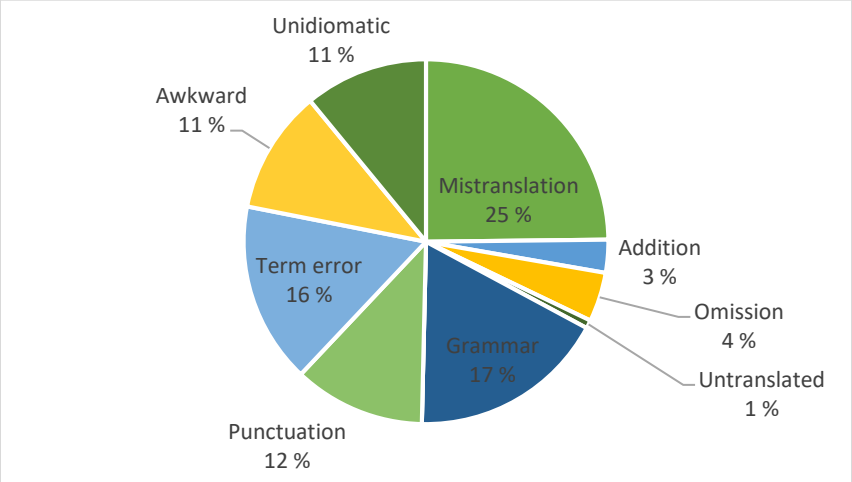



*Figure 4: Error type frequencies for the OD corpus*

of fiction, the engine failed by considering that they belonged to another domain and, as such, the errors were in fact related to domain-specificity. Nevertheless, it can be argued that having a category such as term errors might be an issue if it can potentially only be used with some of the material. Thus, a revised error typology should take this into account and consider whether less ambiguous ways to classify what was now described as mistranslations and term errors could be used. The other error categories were, however, deemed successful and could be used as a basis for further error typologies.

Moving back to the genres, some broad strokes are visible. The Fic corpus had the broadest selection of various errors, scoring the highest numbers in the following categories: awkward, unidiomatic, addition, spelling and untranslated. Here it was perhaps visible that the corpus consisted of language that was more alien to the engines than the EU corpora. EU texts can be considered to follow pretty similar structural patterns, but a more free-flowing fictional text could easily be seen to cause problems for NMT. It should also be noted that the Fic corpus consisted of The Hound of Baskervilles, which is written in somewhat archaic English. This might have been problematic for engines that are expected to have been trained on more modern material. It could also explain the relatively high number of untranslated words and spelling mistakes: out of the entire material, the fic corpus included 67% of all errors regarding of untranslated words and 78% of all incorrectly spelled words. It also made 57% of all errors in the unidiomatic category and 46% in the awkward category.

Awkwardness and unidiomaticity are, however, not necessarily features that indicate issues in understanding the content but that the engine might have had difficulties with the novel's structures. On the other hand, the Fic corpus included 39% of the total number of mistranslations between the three corpora, compared to the PL corpus with 33% and OD with 28%, even though the latter two also made several term errors that were not in the Fic corpus. If the number of term errors and mistranslations are, however, combined, the Fic corpus includes 30% of them and both the PL and OD corpora 35%, meaning that the Fic corpus actually made less mistranslations than the other two. It might be that even though the engine failed to translate some of the rarer and more archaic words, thus the relatively large number of untranslated words and spelling errors, fictional text in general are expected to include more non-specialised content which might actually make them easier to translate. This supports the notion that the engines had more difficulties with the Fic corpus's structures than understanding its content. It would be interesting to see whether the results would be similar had the corpus included some more modern works or if an engine had been used that had been trained solely on fictional material, as was suggested in the previous chapter.

Even though the EU corpora were expected to perform somewhat similarly to each other, there are some differences. The OD corpus, for instance, included more mistakes related to grammar (39%), awkwardness (38%) and unidiomaticity (30%) compared to the PL corpus (23%, 15% and 12%), although both were still greatly behind the Fic corpus. This might be because, according to Biber's dimensions, official documents are expected to be more informational and abstract than professional letters and, as will be noted in chapter 4.2.4, the average sentence length in the OD corpus was in fact higher (31.903) than in the PL (24.877) or Fic corpus (22.652). On the other hand, the PL corpus included the most omissions out of the three (46%) and is, in that sense, much closer to the Fic (39%) than the OD corpus (15%). It would be easy to presume that official documents are "harder" and more complex if they are more informational and abstract, but omissions are generally expected to happen in such places, as in Example 9 below, where the engine has failed to reproduce a small part of the source text. It is, thus, unclear why the PL still includes so many omissions and it would be interesting to study it further to see whether the result is accurate or just a statistical fluke.

| Source text | NMT output | Original translation |
|---|---|---|
| This provision is confirmed by Article 16 of the EC Treaty, concerning services of general economic interest, which was introduced by the Amsterdam Treaty and entered into force on 1 May 1999 - Article 16 states: "**Without prejudice to Articles 73, 86 and 87**, and given the place occupied by services of general economic interest in the shared values of the Union as well as their role in promoting social and territorial cohesion, the Community and the Member States, each within their respective powers and within the scope of application of this Treaty, shall take care that such services operate on the basis of principles and conditions which enable them to fulfil their missions". | Tämä säännös vahvistetaan EY:n perustamissopimuksen 16 artiklassa yleishyödyllisistä taloudellisista palveluista, jotka on otettu käyttöön Amsterdamin sopimuksella ja jotka tulivat voimaan 1 päivänä toukokuuta 1999 - 16 artiklassa todetaan seuraavaa: "Ottaen huomioon yleisiin taloudellisiin tarkoituksiin liittyvien palvelujen harjoittama paikka unionin yhteisissä arvoissa sekä niiden rooli sosiaalisen ja alueellisen yhteenkuuluvuuden edistämisessä yhteisö ja jäsenvaltiot huolehtivat siitä, että ne toimivat näiden periaatteiden mukaisesti ja niiden soveltamisalan mukaisesti." | Tämä määräys vahvistetaan EY:n perustamissopimuksen 16 artiklassa, joka sisällytettiin Amsterdamin sopimuksella ja tuli voimaan 1 päivänä toukokuuta 1999, sekä jossa määrätään seuraavaa: "Ottaen huomioon yleistä taloudellista etua koskevien palvelujen tärkeän aseman unionin yhteisten arvojen joukossa ja niiden merkityksen sosiaalisen ja alueellisen yhteenkuuluvuuden edistämisessä yhteisö ja jäsenvaltiot huolehtivat kukin toimivaltansa mukaisesti ja tämän sopimuksen soveltamisalalla siitä, että tällaiset palvelut toimivat sellaisin perustein ja edellytyksin, että ne voivat täyttää tehtävänsä, sanotun kuitenkaan **rajoittamatta 73, 86 ja 87 artiklan soveltamista**". |

*Example 9: Omission in the OD corpus*

### 4.2.2 Comparing the two engines

Overall, the two engines performed relatively similarly. Numerically, Lingsoft's segments included 182 errors and Google's 223, although it should, once again, be noted that a single segment could include multiple error types and that each occurrence was counted only once per segment. For this part of the analysis, both sets were weighted equally to provide a balanced analysis. The weighted error distribution between the two engines can be seen below in Figure 5.

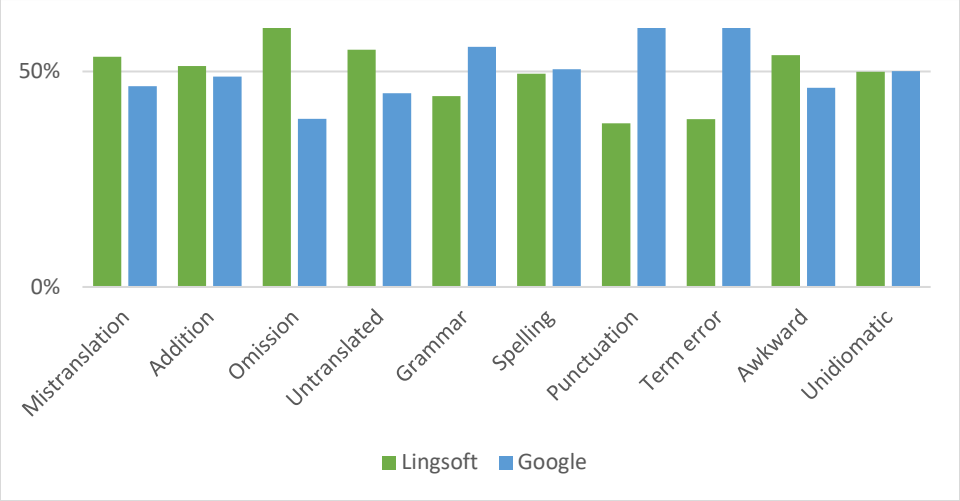



*Figure 5: Equally weighted error distribution between Lingsoft's and Google's NMT engines*

In most categories, the engines are surprisingly close to each other. The only three major differences are on omissions, punctuation and term errors. Interestingly, 48% of Lingsoft's omissions happened in the Fic corpus, compared to Google's 28%. Thus, the high number might be more due to the engine's incapability of processing the Fic corpus, which might in turn raise from the fact that Lingsoft is expected to have had less training material. In any case, making broader conclusions would require more data.

Regarding punctuation, there were two simple reasons that explain why Google made more errors with it. Google's NMT engine, for whatever reason, often added unnecessary whitespaces before and after certain characters (/, <, >), as can be seen in Example 10 below. In want of a better category, these were marked as punctuation errors.

| Source text | NMT output | Original translation |
|---|---|---|
| Commission Decision of 12 April 1999 (IV/M.1482 - Kingfisher/Großlabor, recital 26); | Komission päätös, tehty 12 päivänä huhtikuuta 1999 (IV / M.1482 - Kingfisher / Großlabor, johdanto-osan 26 kappale); | Komission päätös, tehty 12 päivänä huhtikuuta 1999 (IV/M.1482 - Kingfisher/Großlabor, 26 kohta); |

*Example 10: Unnecessary whitespaces in Google's NMT output*

A separate peculiarity to Google's NMT engine was that whenever it appeared to lose its train of thought and was supposedly unable to parse together the end of a long sentence, it just ended it using either a period or a semicolon and just entered whatever was left after it, as can be seen in Example 11. The reason why Google's engine did this is unclear, as Lingsoft's engine was always able to successfully finish its sentences in all the analysed segments. These were also marked as punctuation errors, among other things.

| Source text | NMT output | Original translation |
|---|---|---|
| 3.2. Regular and ad hoc meetings may be held between officials of the two parties with notification to, and participation as far as possible, of liaison officials, covering practical matters of cooperation, in particular the implementation of projects and the participation in committees, groups and working parties and the preparation of documents. | 3.2. Kummankin osapuolen virkamiesten välillä voidaan järjestää säännöllisiä ja tilapäisiä kokouksia, joissa ilmoitetaan yhteyshenkilöille ja osallistutaan niihin mahdollisimman laajasti ja jotka kattavat käytännön yhteistyötä koskevat kysymykset, erityisesti hankkeiden toteuttamisen ja osallistumisen komiteoihin, ryhmiin ja työhön**. osapuolten ja asiakirjojen valmisteluun.** | 3.2 Osapuolten virkamiesten välillä voidaan järjestää säännöllisiä ja erityisiä tapaamisia, joista on ilmoitettava yhteyshenkilöille, joiden olisi mahdollisuuksien mukaan myös osallistuttava tapaamisiin. Kokouksissa käsitellään yhteistyön käytännön kysymyksiä, erityisesti hankkeiden toteuttamista sekä osallistumista komiteoihin, ryhmiin ja työryhmiin ja asiakirjojen laatimista. |

*Example 11: Google fails to parse together entire sentence*

On term errors, it is difficult to say why Google's engine appeared to perform so poorly, although 68% of Google's term errors are in the OD corpus, compared to Lingsoft's 38%. This is peculiar as EU texts are generally broadly available for training purposes and, as such, their terminology should be relatively simple for the engines. One could of course speculate that considering the amount of training data that Google has used on their engine, the engine might no longer be able to prioritise EU material when translating similar material. But once again, more data would be needed to draw broader conclusions.

## 4.3 Sentence length analysis

The sentence length analysis was carried out by calculating the average word count for each segment of the corpus and then comparing it to the word count of the segments used in the segment level analysis. In Table 5, the average sentence length is compared to the system level LeBLEU score averages used in chapter 4.1.2.

| Sentence length | In all | LeBLEU score |
|---|---|---|
| Fic | 22.652 | 0.480 |
| PL | 24.877 | 0.675 |
| OD | 31.903 | 0.666 |

*Table 5: Sentence length in all material vs. LeBLEU score*

Looking at the system level LeBLEU scores, there appears to be little correlation between them and the average sentence length in each corpus. The OD corpus shows an increase of over 40% in average sentence length over the Fic corpus, but it still received much better evaluation scores. Naturally, there is no individual indicator that can be used to predict how well a certain text will be translated but a multitude of separate factors is required. As such, an analysis was needed to compare whether the increase in LeBLEU score in a single corpus was visible in the average sentence length. This analysis of the Fic corpus is presented below in Table 6.

| LeBLEU | Sentence length |
|---|---|
| 0.0 | 2,883 |
| 0.2 | 7,883 |
| 0.4 | 16,517 |
| 0.6 | 15,55 |
| 0.8 | 6,017 |

*Table 6: Relation of LeBLEU score to average sentence length*

The results are somewhat inconclusive as the low end of the scale (0.0-0.2) does not match the rest of the results at all. It is, however, somewhat expected that the segments with the absolute worst scores are somewhat short: if there are only a few words, there are probably relatively few 'filler' words (*if*, *that,* and) that the engine is expected to almost always get right. A short sentence also provides less context for the engine to use in their translation. As such, the engine probably either gets a lot right or a lot wrong, as can be seen in Example 12 below. This does not, however, explain why the segments with the highest score (0.8) were still relatively short.

| Source text | NMT output | Original translation |
|---|---|---|
| Dear me! | Hei! | Kas vaan! |

*Example 12: Short segment with 0.000000 score*

In the case of the Fic corpus, the translator also appears to have used the most creative liberties with the shortest segments. As can be seen in in Example 12 and Example 13 below, where the translator has translated the source text as "Goodbye, dear gentlemen".

| Source text | NMT output | Original translation |
|---|---|---|
| Au revoir, and good-morning! | Kumoa ja aamulla! | Näkemiin, hyvät herrat. |

*Example 13: Creative liberties in short segment*

In the slightly higher end of the scale, there does, however, appear to be some correlation. Starting from 0.4, the average sentence length does decrease with every following increase in score. The drop between 0.6 and 0.8 is, however, remarkably and almost questionably large. Thus, there appears to be some correlation between a lower sentence length and an increase in evaluation scores, but too short sentences can still be highly problematic.

# 5 Conclusions

This study set out to find whether there were discernible features that could be used to predict whether certain texts might be suitable for NMT. As this was a topic with relatively little previous research (Calude 2004, Salimi 2014), ways to measure this objectively had to be discovered. Biber's (1988, 1989) dimensions were used as a basis for choosing texts in three separate genres to see whether the linguistic aspects of a text could be used to predict how translatable it was to NMT. The material was translated using two NMT engines by Google and Lingsoft. The translations were subsequently analysed against reference translations using automatic evaluation, more specifically LeBLEU. A selection of the material was also analysed by hand to see which error types appeared to be the most prevalent and whether there was variance between the three genres and the two engines. A brief sentence length analysis was also carried out to see whether there was correlation between a corpus's evaluation scores and its average sentence length.

Regarding Biber's dimensions, it was discovered in the study that non-narrative texts appeared to be more suitable for NMT than narrative texts and explicit texts more suitable than situation-dependent. On the other hand, only three genres were used in the study and more would be needed to be able to see if and which of the five dimensions could be accurately used to predict how suitable a text is for NMT. It was, however, deemed somewhat difficult to find enough material in some of the genres as a reference translation by a human translator is required for automatic evaluation. Transliterated conversations or interviews would have been a great addition, but no such material was discovered freely available. One could naturally argue that spoken conversations are not expected to be a large focus for NMT research, but on the other hand automatic interpreting applications have been on the rise for the past few years (see e.g. Kohn 2019) so there is a growing market for them as well. In any case, choosing Finnish as the second language for the study might have made finding material slightly more difficult as was presumed, and more might have been available in language pairs such as English-Spanish or English-French.

During the analysis, it was hypothesised that the material chosen for the study might not have been ideal for an accurate study. The Fic corpus consisted solely of The Hound of Baskervilles, a book with relatively archaic language, which might have factored into how well the engines were able to perform. This might have been visible in the amount of errors both engines did with untranslated and incorrectly spelled words in the Fic corpus. In the case of the two EU corpora used in the study, both were chosen

due to how well they fit in to Biber's dimensions, but their free and well-known availability might have made them slightly less ideal for a NMT study. As practically all EU material and their translations are freely available online in simple bilingual formats, it is more than expected that most NMT engines have included them in their training material. As such, the texts and their content might have already been somewhat familiar to both engines, which might have made their scores somewhat higher than if official documents or professional letters from other contexts had been used.

Despite this, the Fic corpus included less mistranslations and term errors than the other two corpora if the two numbers were summed up, although it made the most mistakes related to awkwardness and unidiomaticity. The EU corpora performed relatively similarly, although there were some differences: The OD corpus made much more mistakes related to grammar (39% of all three corpora), awkwardness (38%) and unidiomaticity (30%) compared to the PL corpus (23%, 15% and 12%, respectively), whereas the PL corpus included the most omissions. According to Biber's dimensions, official documents were expected to be more informational and abstract than professional letters, and the OD corpus also had a higher sentence length (31.903) than the PL corpus (24.877). The results were, however, inconclusive on whether this explains the differences and more research would be needed.

Between the error frequencies of the two engines, only three major differences were discovered: with omissions, punctuation and term errors. The difference in omissions was due to Lingsoft's engine's difficulties with the Fic corpus, which was speculated to be due to its lower amount of training material compared to Google. On punctuation, Google was discovered to have made certain errors that Lingsoft never did, which lead to a notable increase in the category. Google also made much more term errors than Lingsoft with the EU corpora, and it was speculated that the expected greatness of Google's training data might have, in turn, lead to it being slightly poorer in dealing with domain-specific terminology. The answer to this might be domain adaptation, which was mentioned in chapter 2.2.2.

An interesting result was also that Google's Cloud Translate was better than Lingsoft's NMT engine in all three of the genres. It had been hypothesised that Lingsoft might be able to perform better since their engine had been developed in Finland for Finnish specifically, but this did not reflect in the evaluation scores. This might indicate that the amount of training data is key with NMT, and the fact that neural networks are able to learn autonomously which features and classifiers they should be looking at in a data set (Maturana & Scherer 2015: 1) might make understanding the features of a language less important than with PBSMT, for instance. On the other hand, what is going on under the hood of a NMT

engine should not be disregarded, as for instance new encoder-decoder models were speculated to have led to an increase in quality with longer sentences, as was mentioned in chapter 2.2.2. It is, arguably, also possible that Lingsoft's engine did in fact gain an advantage from its knowledge of Finnish but that Google's engine was much better in so many other fronts. Proving this objectively is, however, impossible without knowing the specific setup of each engine or making a comparative study between them using the same amount of training data.

A side result in the study was that both NMT and automatic evaluation appear to be biased against the translations of fiction and other genres that do not conform to the same rules as the material that is usually used to train NMT engines. Training material is generally expected to be more akin to the EU corpora with relatively normative sentence structures and where the translations are always expected to follow the source text closely. It would, thus, be interesting to train an NMT engine purely with fictional texts and see how big an impact it would make to the 'creativity' of the engine and whether that might lead to less accurate translations. Finding modern material and their reference translation might, however, once again prove difficult, this time also for reasons related to copyrights.

In the case of evaluating translations of fictional texts, it was suggested that comparing NMT output to reference translations is problematic, as a translation of a fictional text is expected to take some degree of creative liberty with the source text, as was evident in the closer analysis of the Fic corpus. As the evaluation metric is unable to understand situations where the NMT output might in fact be much closer to the source text, it will give a low score to any translation that does not appear to be similar to the surface meaning of a reference translation. Thus, "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al. 2002: 1) clearly does not tell the whole story.

In chapter 3.2.2 on automatic evaluation, it was also discussed that evaluation metrics themselves should be recalibrated whenever things such as a target language or MT type changes, as the metric will always be biased towards material that is similar to the one it has been trained with (Babych 2014: 468-469). I would also argue that the need for recalibration should be extended to the material's genre or text type as well and that a metric should be developed that takes the semantic meaning of the source, target and reference translation into account instead of purely looking at its surface value. In its current form, it is difficult to see how automatic evaluation could be considered a good way to measure the translation quality of fictional or other unusual texts compared to those that NMT and automatic evaluation is traditionally used on.

The final part of the analysis regarding average sentence length was inconclusive. It was obvious from the results that sentence length cannot be used to predict how an individual text or data set might translate, as the corpus with the longest average sentence length (OD) received much better evaluation scores than the one with the shortest (Fic). Within a single data set, there might, however, be some correlation. The closer analysis of the Fic corpus showed that segments with very low scores (0.0-0.2) were often exceedingly short. This was speculated to be caused by the fact that if a segment is short, the engine is expected to either get a lot wrong or a lot right as opposed to longer segments, where the engine might make some mistakes but still be able to work with other parts of the segment, which is expected to lead to a score with greater variability in the middle of the scale. Segments with higher scores (0.4-0.8), however, appeared to show some correlation and there was a relatively steep dip between the average sentence length of the 0.6 set (15,55) and the 0.8 set (6,017). This appears to show that, within a single data set, shorter segments appear to translate better on average but too short ones are potentially still problematic, as was evident in the 0.0 set (2,883).

Overall, the methods used in this study could easily be used in future studies and expanding them would be relatively simple. Apart from the manual error categorisation, increasing the number of data sets, languages, NMT engines or evaluation metrics would not translate to an equal increase in the amount of time it would take to process them, but be rather trivial. Filling out the empty slots in Biber's dimensions would be a simple and obvious continuation of this study. Finding enough relative material was, however, deemed somewhat difficult, although the situation might be better with other language pairs.

The error categorisation adapted from the DQF-MQM typology (TAUS 2019) was mostly successful, although some changes should be made to the way mistranslations and term errors are classified. Based on this study, I propose expanding the concept of mistranslations and term errors into three separate categories: 1) wrong grammatical relations, 2) wrong domain or context and 3) mistranslation with no obvious reason. This should make it easier to compare different types of mistranslations, as an NMT engine translating the subject of a sentence as its object is expected to have been caused by an entirely separate issue than if the same engine uses a valid synonym for a term that is wrong in the current context. Identifying mistranslations of words with the wrong context or domain would also avoid the issue of determining what is a term and whether all texts necessarily include them. An additional category should also be retained for mistranslations where it is impossible to say why and where an engine has failed, or if it has made an unforced error by translating "cucumber" as "habit", for instance.

In conclusion, it is obvious that studying the suitability of different texts for NMT, or MT in general, is a multifaceted issue and that there is no one single method that can be used to provide a generalisation that works in all situations. A good example of this was average sentence length that showed that even though there was some correlation between the scores of individual segments and their average sentence length in a single data set, it was still only one of multiple factors and could not be used to predict the translatability of texts against each other. In any case, it is obvious that linguistic features can be used to predict whether a text is suitable for NMT. As such, more research with more material is needed and the methods proposed in this study are a good first step towards that.

# References

ALPAC (1966): *Language and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee*. Washington, DC: National Academy of Sciences. Available online: http://www.mt-archive.info/ALPAC-1966.pdf (Accessed 11.12.2018)

Arthur, Philip; Neubig, Graham & Nakamura, Satoshi (2016): Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. 1557–1567. Available online: https://arxiv.org/pdf/1606.02006.pdf (Accessed 10.12.2018)

Atkins, Sue; Clear, Jeremy & Ostler, Nicholas (1992): Corpus design criteria. *Literary and linguistic computing* 7.1. Available online: http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf (Accessed 26.11.2018)

Bahdanau, Dzmitry; Cho, Kyunghyun & Bengio, Yoshua (2014): Neural machine translation by jointly learning to align and translate. Available online: https://arxiv.org/abs/1409.0473 (Accessed 28.11.2018)

Bentivogli Luisa, Bisazza Arianna, Cettolo Mauro, Federico Marcello (2016): Neural versus phrase-based machine translation quality: a case study. In: *Proceedings of the 2016 conference on empirical methods in natural language processing* (EMNLP 2016), Austin, Texas. 257–267. Available online: https://arxiv.org/pdf/1608.04631.pdf (Accessed 13.12.2018)

Bentivogli Luisa, Bisazza Arianna, Cettolo Mauro, Federico Marcello (2018): Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. *Computer Speech & Lang* 49: 52-70.

Biber, Douglas (1988): *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas (1989): A typology of English texts. *Linguistics*. 27. 3-44. Available online: https://www.degruyter.com/view/j/ling.2013.51.issue-jubilee/ling-2013-0040ad.pdf (Accessed 26.11.2018)

Biber, Douglas (1993): Representativeness in corpus design. *Literary and linguistic computing*, 8(4), 243-257. Available online: http://otipl.philol.msu.ru/media/biber930.pdf (Accessed 26.11.2018)

Brants, Thorsten; Popat, Ashok C.; Xu, Peng; Och, Franz J. & Dean, Jeffrey (2007): Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 858-867. EMNLP-CoNLL. Available online: http://www.aclweb.org/anthology/D07-1090 (Accessed 28.11.2018)

Bühler, Karl (1934): *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Stuttgart. Gustav Fischer.

Calude, Andreea (2004): Machine Translation of Various Text Genres. Te Reo–the New Zealand Linguistic Society Journal, 46, 67-94. Available online: https://www.researchgate.net/publication/228938192_Machine_translation_of_various_text_genres (Accessed 31.10.2018)

Crego, Josep; Kim, Jungi; Klein, Guillaume; Rebollo, Anabel; Yang, Kathy; Senellart, Jean; Akhanov, Egor; Brunelle, Patrice; Coquard, Aurélien; Deng, Yongchao; Enoue, Satoshi; Geiss, Chiyo; Johanson, Joshua; Khalsa, Ardas; Khiari, Raoum; Ko, Byeongil; Kobus, Catherine; Lorieux, Jean; Martins, Leidiana; Nguyen, Dang-Chuan; Priori, Alexandra; Riccardi, Thomas; Segal, Natalia; Servan, Cristophe; Tiquet, Cyril; Wang, Bo; Yang, Jin; Zhang, Dakun; Zhou, Jing; Zoldan, Peter (2016): SYSTRAN's Pure Neural Machine Translation Systems. Available online: https://arxiv.org/pdf/1610.05540.pdf (Accessed 11.12.2018)

Fawcett, Peter (1997): *Translation and Language: Linguistic Theories Explained*. Manchester: St. Jerome.

Federmann, Christian (2018): Appraise Evaluation Framework for Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations (pp. 86-88). Available online: https://aclweb.org/anthology/C18-2019 (Accessed 13.4.2019)

Forcada, Mikel L. (2017): Making sense of neural machine translation. *Translation Spaces* 6.2: 291-309. Available online: https://www.dlsi.ua.es/~mlf/docum/forcada17j2.pdf (Accessed 27.11.2018)

Gehring, Jonas; Auli, Michael; Grangier, David; Yarats, Denis & Dauphin, Yann N. (2017): Convolutional Sequence to Sequence Learning. eprint arXiv:1705.03122. Available online: https://arxiv.org/abs/1705.03122 (Accessed 28.11.2018)

Goldberg, Yoav (2017): *Neural network methods for natural language processing*. Synthesis Lectures on Human Language Technologies 10.1. San Rafael, California: Morgan & Claypool.

Heikkinen, Vesa; Voutilainen, Heikki; Lauerma, Petri; Tiililä, Ulla & Lounela, Mikko (2012): *Genreanalyysi: Tekstilajitutkimuksen käsikirja*. Gaudeamus, Helsinki.

Homem, Nuno & Carvalho, Joao Paulo (2011): Authorship identification and author fuzzy fingerprints. In *Proc. of the NAFIPS2011-30th Annual Conference of the North American Fuzzy Information Processing Society*.

Hutchins, John (2007): Machine translation: A concise history. In *Computer aided translation: Theory and practice* 13: 29-70. Available online: http://www.hutchinsweb.me.uk/CUHK-2006.pdf (Accessed 10.12.2018)

Karlgren, Jussi & Cutting, Douglass (1994): Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING '94): 1,071–1,075. Available online: http://www.aclweb.org/anthology/C94-2174 (Accessed 27.11.2018)

Kessler, Brett; Nunberg, Geoffrey & Schütze, Hinrich (1997): Automatic detection of text genre. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics:* 32-38. Association for Computational Linguistics. Available online: http://www.aclweb.org/anthology/P97-1005 (Accessed 27.11.2018)

Koehn, Philipp & Knowles, Rebecca (2017): Six Challenges for Neural Machine Translation. In: *Proceedings of the First Workshop on Neural Machine Translation:* 28–39 Available online: https://arxiv.org/pdf/1706.03872.pdf (Accessed 7.12.2018)

Kohn, Marek (2019): Is the era of artificial speech translation upon us. In *The Guardian*. Abridged from *Four Words for Friend: Why Using More Than One Language Matters Now More Than Ever*. Yale

University Press, New Haven. Available online: https://www.theguardian.com/technology/2019/feb/17/is-the-era-of-artificial-speech-translation-upon-us (Accessed 13.4.2019)

Lee, David Y. W. (2001): Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle. University of Wollong. Available online: https://ro.uow.edu.au/artspapers/598/ (Accessed 20.11.2018)

Li, Haoxiang; Lin, Zhe; Shen, Xiaohui: Brandt, Jonathan & Hua, Gang (2015): A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 5325-5334). Available online: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Li_A_Convolutional_Neural_2015_CVPR_paper.pdf (Accessed 17.12.2018)

Luong, Minh-Thang & Manning, Christopher D. (2015): "Stanford neural machine translation systems for spoken language domains." *Proceedings of the International Workshop on Spoken Language Translation. Available online*: https://nlp.stanford.edu/pubs/luong-manning-iwslt15.pdf (Accessed 10.12.2018)

Luong, Minh-Thang, Ilya Sutskever, Quoc V. Le, Oriol Vinyals & Wojciech Zaremba (2014): Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*. Available online: http://aclweb.org/anthology/P15-1002 (Accessed 14.12.2018)

Maturana, Daniel & Scherer, Sebastian (2015): Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*: 922-928). IEEE. Available online: https://www.ri.cmu.edu/pub_files/2015/9/voxnet_maturana_scherer_iros15.pdf (Accessed 17.12.2018)

Munday, Jeremy (2013): *Introducing translation studies: Theories and applications*. 3rd edition. London, Routledge.

Nord, Christiane (2005): *Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis*. 2nd edition. Amsterdam: Rodopi.

Pietikäinen, Sari & Mäntynen, Anne (2009): *Kurssi kohti diskurssia*. Tampere: Vastapaino.

Reiss, Katharina (1971): *Möglichkeiten und Grenzen der Übersetzungskritik*. Munich, Max Hueber.

Reiss, Katharina (1977): Text types, translation types and translation assessment. Translated by Andrew Chesterman. In *Readings in Translation Theory*: 105-115. Finn Lectura, Helsinki, 1989.

Salimi, Jonni (2014): Machine Translation of Fictional and Non-fictional texts. An examination of Google Translate's accuracy on translation of fictional versus non-fictional texts. Bachelor's Degree Project. Stockholms Universitet. Available online: http://www.diva-portal.org/smash/get/diva2:737887/FULLTEXT01.pdf (Accessed 31.10.2018)

Sennrich, Rico; Haddow, Barry & Birch, Alexandra. (2015): Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*. Available online: http://www.aclweb.org/anthology/P16-1162 (Accessed 7.12.2018)

Shore, Susanna & Mäntynen, Anne (2006): Johdanto. In *Genre – tekstilaji* (Vol. 213). Mäntynen, Anna, Shore, Susanna, & Solin, Anna (Eds.). Suomalaisen Kirjallisuuden Seura. P. 9-41

Stamatatos, Efstathios; Fakotakis, Nikos & Kokkinakis, George (2000): Automatic text categorization in terms of genre and author. In *Computational linguistics* 26.4: 471-495.

Sutskever, Ilya; Vinyals, Oriol & Le, Quoc V. (2014): Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, edited by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, Kilian Q. Weinberger: 3104-3112. Available online: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf (Accessed 28.11.2018)

Toral, Antonio & Sánchez-Cartagena, Víctor M. (2017): "A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions." arXiv preprint arXiv:1701.02901. Available online: https://arxiv.org/abs/1701.02901 (Accessed 14.12.2018)

Vaswani, Ashish; Shazeer, Noam; Niki, Parmar; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz & Polosukhin, Illia (2017): Attention is all you need. eprint arXiv:1706.03762. Available online: https://arxiv.org/abs/1706.03762 (Accessed 28.11.2018)

Werlich, Egon (1976): A text grammar of English. Quelle & Meyer.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouz (2016): Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Available online: https://arxiv.org/pdf/1609.08144.pdf (Accessed 31.10.2018)

# Appendix 1: Lyhennelmä

Helsingin yliopisto

Humanistinen tiedekunta

Englannin kääntäminen

Ari Gröhn: Neuroverkkokonekääntämisen soveltuvuus erityyppisille teksteille: tutkimus potentiaalisista indikaattoreista

Pro gradu -tutkielma, 47 s., suomenkielinen lyhennelmä 10 s.

Huhtikuu 2019

## 1 Johdanto

Konekääntäminen ei ole keksintönä uusi, vaan erilaisia käänninmalleja on esitetty 1940-luvulta lähtien (Hutchins 2007). Uutuutena on kuitenkin neuroverkkoihin ja syväoppimiseen perustuva neuroverkkokonekääntäminen (Forcada 2017), jonka nimeen suuret toimijat kuten Google (Wu et al. 2016) ovat vannoneet viimeisten vuosien ajan. Neuroverkkokonekäänninten on havaittu lisäävän merkittävästi konekäännösten sujuvuutta, mutta niillä on välillä ongelmia tarkkuuden kanssa (ks. luku 2.2).

Tämän tutkielman tarkoituksena on selvittää, onko olemassa kielellisiä ominaisuuksia, joita voidaan käyttää ennustamaan, onko jokin teksti soveltuva tai soveltumaton neuroverkkokonekäännettäväksi. Näiden ominaisuuksien perusteella olisi mahdollista päätellä, millaisia tekstejä kannattaisi konekääntää neuroverkkokääntimillä ja millaisia ei. Koska tällaista tutkimusta ei ole kattavasti neuroverkkokonekääntimille vielä tehty, ei olemassa ollut valmiita tutkimusmetodeja. Tämän tutkielman osatavoitteena onkin esittää erilaisia metodeja, joita voisi hyödyntää jatkotutkimuksessa.

Tutkielman analyysissä käytetään kolmea tekstikorpusta, jotka on koottu Biberin (1988, 1989) tekstityyppiluokittelun perusteella kolmesta eri genrestä: fiktiosta, virallisista kirjeistä ja virallisista dokumenteista. Kukin korpus koostuu alkuperäisestä englanninkielisestä lähtötekstistä, suomenkielisestä ihmisen tekemästä referenssikäännöksestä sekä kahden neuroverkkokonekääntimen käännöksestä. Korpukset evaluoidaan automaattisesti käyttäen LeBLEU:ta, joka vertaa referenssikäännöksiä konekäännöksiin ja arvioi yksittäisten lauseiden sekä koko korpusaineiston käännösten onnistumista. Jokaisesta korpuksesta otetaan myös pieni otos, jolle tehdään tarkempi manuaalinen virheluokittelu, jossa tarkastellaan kunkin korpuksen virhejakaumaa ja verrataan niitä

toisiinsa sekä kahden kääntimen välillä. Lopuksi tarkastellaan, löytyykö korpusten keskiarvoisten lausepituuksien ja niiden saamien evaluaatiotulosten väliltä korrelaatiota.

Luku 2 tarkastelee soveltuvia tekstityyppiluokitteluja ja neuroverkkokonekääntämistä. Luku 3 esittelee tutkielmassa käytetyn materiaalin, konekääntimet sekä automaattisen evaluaatioinnin. Luku 4 käsittelee korpustason analyysituloksia sekä virheluokittelun ja lausepituusanalyysin havaintoja. Luku 5 vetää aiemman asian yhteen sekä tekee yleistyksiä ja ehdotuksia potentiaalista jatkotutkimusta ajatellen.

# 2 Teoriatausta

Tämä luku esittelee erilaisia tekstityyppiluokitteluja luvussa 2.1 ja luvussa 2.2 neuroverkkokonekääntämistä, neuroverkkokäänninten tyypillistä rakennetta ja käännöksissä esiintyviä tyypillisimpiä virheitä.

## 2.1 Tekstityypit

Tekstityyppi ei käsitteenä ole aivan yksiselitteinen, vaan useat tutkijat ovat käyttäneet samaa termiä kuvaamaan erilaisia asioita. Esimerkiksi Reiss (1977) luokittelee tekstityyppejä tekstien funktioiden mukaan, kun taas Werlichin (1976) mielestä niillä on paras kuvata tekstien kontekstuaalista fokusta. Oman ongelmansa tuovat myös eri koulukuntien väliset erot: Saksassa lingvistit tekevät eron tekstin funktionaalisen luokittelun (esim. informatiivinen, ekspressiivinen) ja kontekstuaalisen luokittelun (artikkeli, kirja), kun taas englanninkieliset lingvistit käyttävät tekstityyppiä toisinaan kuvaamaan molempia (Nord 2005: 20). Saksalaisen lingvistiikan kaltainen eronteko on nähtävissä myös Suomessa, jossa tekstityypillä ja genrellä on selvä ero (Shore & Mäntynen 2006). Tässä tutkielmassa tekstityypillä viitataankin suomalaisen lingvistiikan tavoin tekstin kielellisiin ominaisuuksiin ja genrellä tarkoitetaan kontekstuaalisia piirteitä.

Tekstityyppiluokittelut voidaan jakaa karkeasti kahteen luokkaan: kvalitatiivisiin ja kvantitatiivisiin. Kuuluisimpia kvalitatiivisia luokitteluja ovat Egon Werlichin (1976) ja Katharina Reissin (1977) esittämät luokittelutavat, jotka eroavat toisistaan merkittävästi. Werlich jakaa tekstit viiteen kategoriaan sen mukaan, millaisen strategian tekstin luoja valitsee: deskriptiiviseen, narratiiviseen, ekspositoriseen, argumentatiiviseen ja instruktiiviseen (Werlich 1976: 39–41). Kullakin tekstityypillä on sille ominaisia kielellisiä elementtejä, ja esimerkiksi deskriptiiviset tekstit sisältävät paljon havaitsemiseen ja olemiseen liittyviä adjektiiveja ja verbejä (ibid.). Reissin (1977) kolme tekstityyppiä taas perustuvat kunkin tekstin kommunikatiiviseen tavoitteeseen: informatiivisuuteen, ekspressiivisyyteen tai operatiivisuuteen.

Kvantitatiivisista tekstityyppiluokitteluista tunnetuin on Biberin (1988, 1989) korpusanalyysiin perustuva luokittelu. Tutkimuksessaan Biber tutkii faktorianalyysillä 23 eri genreä kahdessa eri korpuksessa ja jakaa ne tulostensa perusteella viiteen tyyppiin eli *dimensioon*, joista kullakin on omat tyypilliset kielelliset piirteensä: 1. Involved vs. informational production, 2. Narrative vs. nonnarrative concerns, 3. Explicit vs. situation-dependent reference, 4. Over expression of persuasion ja 5. Abstract vs. nonabstract style. Biberin dimensiot eivät ole binäärisiä, vaan kukin teksti sijoittuu samanaikaisesti kaikkiin viiteen dimensioon ja saa kunkin ääripäiden välissä erilaisia arvoja. Esimerkiksi kasvotusten käydyt keskustelut sijoittuvat ensimmäisessä dimensiossa aivan yläpäähän (involved), kun taas kolmannessa dimensiossa ne ovat suhteellisen alhaalla (situation-dependent). (Biber 1988, 1989.) Tutkielmassa käytetään Biberin dimensioita, koska skalaarisuuden koettiin vastaavan parhaiten todellisuutta binääristen tekstityyppien sijaan ja koska luokittelu antaa hyvin tarkat prototyyppikuvaukset kunkin dimension ääripään tyypillisistä kielellisistä elementeistä. Biberin dimensioista oli myös helppo valita erilaisia genrejä tutkimuksen analyysimateriaaliksi, koska hän kuvaa tutkimuksessaan kaikkien 23 genren sijoittumisen kullekin dimensiolle (Biber 1989).

## 2.2 Neuroverkkokonekääntäminen

Neuroverkkokonekääntäminen seuraa aiempien konekäänninmallien (esim. tilastollinen ja fraasipohjainen) jalanjäljissä. Neuroverkkokääntimet perustuvat valtaviin monikielisiin korpuksiin, jotka muodostuvat lähtökielisistä teksteistä ja niiden kohdistetuista käännöksistä. Suurin innovaatio neuroverkkokääntimissä on juuri neuroverkko, joka on matemaattinen tiedonkäsittelymalli. Verkko koostuu tuhansista yksittäisistä ”neuroneista”, jotka stimuloivat toisiaan, kun niille syötetään koulutusmateriaalia. Riittävällä datamäärällä neuronit pystyvät arvioimaan ratkaisuja hyvin monimutkaisiin ongelmiin, jotka voivat olla ihmiselle itsestään selviä, mutta koneelle hyvin haasteellisia. (Forcada 2017: 2–4.) Neuroverkkojen ydininnovaatio onkin, että ne pystyvät omatoimisesti oppimaan datasta sen piirteitä ja luokittelua omatoimisesti (Maturana & Scherer 2015: 1), eikä niille esimerkiksi tarvitse erikseen kertoa, mikä lauseessa on subjekti tai mitä koko käsite ylipäätään tarkoittaa.

Neuroverkkokäänninten optimaalinen rakenne on kiivaan tutkimuksen kohteena, ja eri käänninten välillä voi olla suuriakin eroja. Lähtökohtaisesti jokaisessa kääntimessä on kuitenkin erillinen enkooderi ja dekooderi, joista ensimmäinen käy läpi lähtötekstiä ja jälkimmäinen tuottaa itse käännöksen. Tapoja molempien toimintaan on kuitenkin lukuisia (Forcada 2017: 7–8.) ja myös esimerkiksi kääntimen koulutusmateriaalin määrällä, laadulla, esikäsittelyllä ja syöttämistavalla on merkitystä (Koehn & Knowles 2017; Brants et al. 2007). Tämä tekee erilaisten käänninten vertailusta haastellista.

Neuroverkkokäänninten on havaittu tuottavan hyvin paljon sujuvampaa tekstiä kuin aiemmat käänninmallit ja ylipäätään tekevän lähes kaikissa virhekategorioissa vähemmän virheitä (Bentivogli et al. 2016). Käänninmallilla on kuitenkin omat ongelmansa: Neuroverkkokääntimet edellyttävät massiivista koulutusmateriaalimäärää ja voivat suoltaa täysiä käsittämättömyyksiä, jos niille syötetään tekstejä, jotka eivät vastaa niiden koulutuksessa käytettyä materiaalia (Koehn & Knowles 2017: 1–2, 10). Vaikka neuroverkkokäännös voi siis olla hyvin sujuvan näköistä, voi se sisältää semanttisia virheitä, joita voi olla vaikea havaita muuten onnistuneen näköisestä käännöksestä (ibid.). Neuroverkkokäänninten on myös havaittu kompastelevan liian pitkien lauseiden (ibid.) ja tuntemattomien sanojen kanssa (esim. Sutskever et al. 2014; Bahdanau et al. 2014; Wu et al. 2016; Bentivogli et al. 2018), joista jälkimmäinen tosin on vain toinen esimerkki koulutusmateriaaliin kuulumattomasta sisällöstä.

# 3 Materiaali ja metodit

Tämä kappale esittelee tutkimuksessa käytettyä materiaalia ja metodeja. Luku 3.1 keskittyy materiaalivalinnan suhteutumiseen Biberin dimensioihin, kun taas luku 3.2 käsittelee analyysissä käytettyä virheluokittelua ja automaattista evaluaatiota. Lopuksi luku 3.3 esittelee lyhyesti tutkimuksessa käytetyt konekääntimet.

## 3.1 Materiaali

Biberin (1988, 1989) dimensioiden perusteella tutkimukseen valittiin kolme eri genreä: fiktio, viralliset kirjeet ja viralliset dokumentit. Näillä genreillä on mahdollista kattaa laajasti kukin viidestä dimensiosta, joskin tutkimusta olisi myös tukenut jonkin puhekielisemmän korpuksen mukaan ottaminen. Kyseisen kaltaista materiaalia ei kuitenkaan ollut saatavilla tutkielman aikarajoitteiden sisällä, joten sellaisen sisällyttäminen ei ollut tämän tutkimuksen puitteissa mahdollista. Tämän seurauksena ensimmäisen dimension yläosa ja kolmannen sekä neljännen dimension alaosat jäävät aliedustetuiksi.

Korpukset koottiin seuraavasti: Viralliset kirjeet ja dokumentit koottiin EUR-Lexistä, joka on Euroopan unionin kokoama datapankki erilaisia asiakirjoja ja niiden kohdistettuja käännöksiä. Kyseiset korpukset koostuvat 27 virallisesta kirjeenvaihdosta ja 20 virallisesta dokumentista. Fiktio-korpus on OPUS-korpuksen osakorpus (ks. Tiedemann 2012) ja koostuu Baskervillen koiran alkuperäisteoksesta ja sen suomenkielisestä käännöksestä vuodelta 1904.

### 3.2 Metodit

Tutkimuksessa käytettävän virheluokittelun pohjana on DQF-MQM:n virhekategorisointi (TAUS 2019), josta karsittiin pois turhia luokkia. Lopullisessa luokittelussa on yhdeksän eri virheluokkaa: addition, omission, mistranslation, untranslated, grammar, spelling, punctuation, awkward ja unidiomatic. Näiden katsottiin kattavan tarpeeksi hyvin lähdekirjallisuudesta löytyneet tyypilliset ongelmakohdat.

LeBLEU (Virpioja & Grönroos 2015) valikoitui tutkimuksen automaattiseksi evaluaatiomenetelmäksi. Tutkimusalan yleisin evaluaatiomenetelmä, BLEU (Papineni et al. 2002), havaittiin kirjallisuuskatsauksen aikana puutteelliseksi neuroverkkokonekäänninten tulosten arviointiin (ks. esim. Callison-Burch et al. 2006; Shterionov et al. 2018). LeBLEU:n etuna on n-grammien osumien arviointi sumeasti ja vahva korreloivuus ihmisarvion kanssa jo lausetasolla, mikä oli edellytys segmenttitason analyysille.

### 3.3 Tutkimuksen neuroverkkokonekääntimet

Tutkimuksen materiaali käännettiin kahdella neuroverkkokonekääntimellä: Googlen kaupallisella Cloud Translation -kääntimellä ja suomalaisen Lingsoftin kääntimellä. Kummankaan kääntimen rakenteesta ei ole saatavilla tarkkaa dokumentaatiota kaupallisuuden ja yrityssalaisuuksien vuoksi, joskin tiedossa on, että Googlen käännin perustuu heidän omaan järjestelmäänsä (Wu et al. 2016), kun taas Lingsoftin käännin perustuu avoimen lähdekoodin Marianiin. Tutkimuksessa käytettiin kahta erillistä käännintä, koska eriävien rakenteiden ja arkkitehtuurien ajateltiin potentiaalisesti vaikuttavan kunkin kääntimen tyypillisiin virheiseen, jolloin kahden erilaisen kääntimen tulisi tasoittaa toisiaan. Toiseksi kääntimeksi päätyi Lingsoftin käännin, koska Suomessa rakennetun kääntimen ajateltiin toimivan pienen kielialueen suomen kanssa paremmin kuin yleiskääntimenä toimiva Googlen Cloud Translation.

## 4 Tutkimuksen tulokset

Tämä luku käsittelee tutkimuksen tuloksia: luku 4.1 keskittyy korpustason analyysiin, luku 4.2 segmenttitason analyysiin ja luku 4.3 lausepituusanalyysiin.

### 4.1 Korpustason analyysi

Korpustason analyysin tulokset on esitetty alla taulukossa 1.

| Korpus | NMT-käännin | LeBLEU-arvo |
|---|---|---|
| **Fiktio** | Google | 0.507769 |
| | Lingsoft | 0.452538 |
| **Viralliset kirjeet** | Google | 0.688428 |
| | Lingsoft | 0.661735 |
| **Viralliset dokumentit** | Google | 0.683341 |
| | Lingsoft | 0.648748 |

*Taulukko 1: Korpustason LeBLEU-arvot*

Tulosten valossa on ilmeistä, että fiktiokorpus oli kaikista haasteellisin molemmille kääntimille, kun taas EU-materiaalista kootut kaksi muuta korpusta saivat huomattavan paljon parempia tuloksia. Tässä suhteessa tutkimuksen lähtöasetelmissa on kuitenkin voinut olla ongelma kahdesta syystä: EU-materiaalit ovat suosittuja koulutusmateriaaleja konekääntimille, koska ne ovat vapaasti ladattavissa. EU-korpukset ovat myös kielellisesti paljon lähempänä sellaisia tekstejä, joita konekääntimillä tyypillisesti käännetään ja jollaisilla myös evaluaatiomenetelmiä kalibroidaan. Fiktiivinen, narratiivinen materiaali on siis lähtökohtaisestikin ollut tuntemattomampaa sekä kääntimille että evaluaatiomenetelmille, mikä varmasti näkyy myös korpuksen käännöksissä sekä sen saamissa evaluaatiotuloksissa. Fiktiivisen materiaalin vertaaminen referenssikäännöksiin on myös ongelmallista, koska segmenttitason analyysissä kävi hyvin ilmi, että ihmiskääntäjä oli hyvin usein päätynyt käännöksiin, jotka erosivat merkittävästi lähtötekstistä. Vaikka konekäännin olisikin tuottanut semanttisesti täysin lähtötekstiä vastaavan käännöksen, on se voinut saada huonon arvon, koska referenssikäännös onkin päätynyt esimerkiksi lisäämään segmenttiin jotain kerronnallisista syistä. EU-tekstejä päinvastoin taas käännetään hyvinkin kirjaimellisesti ja esimerkiksi lauserakenteet pysyvät hyvin usein täysin vastaavina, joten myös neuroverkkokääntimet ovat oppineet jäljittelemään vastaavia käännösstrategioita. Olisikin mielenkiintoista tutkia neuroverkkokäännintä, joka on koulutettu yksinomaan fiktiivisellä materiaalilla ja käyttää myös evaluaatiomenetelmää, joka on kalibroitu vastaavalla aineistolla.

Googlen tulokset olivat jokaisen korpuksen osalta paremmat kuin Lingsoftin, joten tulosten perusteella näyttäisi, että koulutusmateriaalin määrä ja rakenteelliset ratkaisut ovat neuroverkkokonekääntimillä tärkeämpiä kuin yksittäisen kielen erikoistuntemus. Mahdollista on tietysti myös se, että Lingsoftin käännin itse asiassa hyötyi kieliparista, mutta Googlen käännin oli niin monella muulla saralla parempi. Asiaa on kuitenkin vaikea todistaa ilman tarkempia tutkimuksia, jotka eivät kaupallisten kääntimien osalta ole mahdollisia.

| | Dimensio 1: Involved vs. Informational | Dimensio 2: Narrative vs. Non-narrative | Dimensio 3: Explicit vs. Situation-dependent | Dimensio 4: Overt persuasion | Dimensio 5: Abstract vs. Non-abstract |
|---|---|---|---|---|---|
| Yläryhmä | - | 0.480 | 0.666, 0.678 | 0.678 | 0.666 |
| Keskiryhmä | 0.480, 0.678 | - | 0.480 | 0.480, 0.666 | 0.678 |
| Alaryhmä | 0.666 | 0.678, 0.666 | - | - | 0.480 |

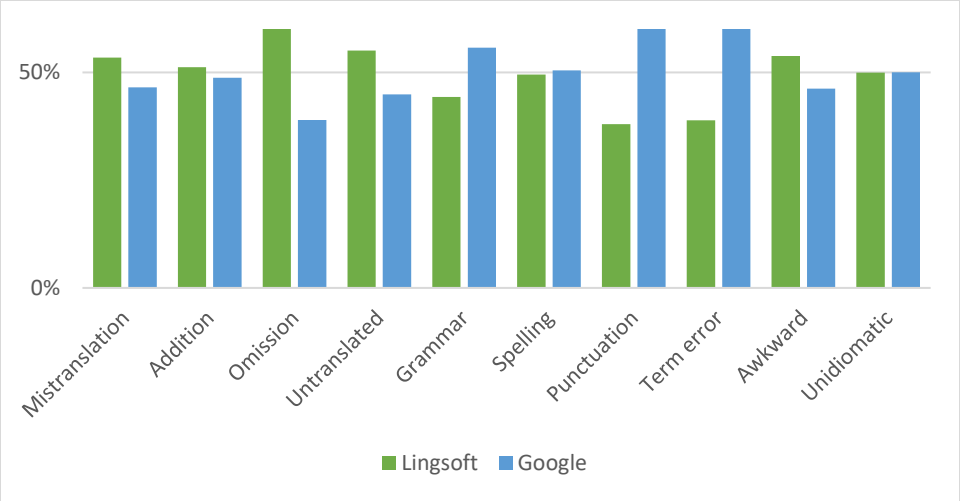
*Taulukko 2: Tutkimuksen genrejen evaluaatiotulokset Biberin dimensioissa*

Biberin dimensioiden osalta tietyissä dimensioissa on havaittavissa korrelaatiota tutkimuksessa käytettyjen kolmen genren perusteella, niin kuin taulukosta 2 näkee. Dimensiot 2 ja 3 näyttäisivät korreloivan jokseenkin vahvasti tulosten kanssa: narratiiviset tekstit näyttäisivät kääntyvän huonommin kuin ei-narratiiviset ja tilannekohtaiset tekstit huonommin kuin eksplisiittiset. Dimensioissa 1, 4 ja 5 on havaittavissa jonkin verran korrelaatiota, mutta tulokset eivät ole niin selviä kuin dimensioissa 2 ja 3. Tulosten varmistaminen kaikkien osalta vaatisikin enemmän analysoituja genrejä ja taulukon nykyisten aukkojen täyttämistä.

## 4.2 Segmenttitason analyysi

Segmenttitason tulokset tukevat ajatusta, että fiktiokorpuksessa käytetty materiaali oli kääntimille kaikista vierain ja polveilevin. Korpus sisälsi eniten virheitä useissa eri virheluokissa (*awkward*, *unidiomatic*, *addition*, *spelling* ja *untranslated*) ja huomionarvoisesti koko analysoidusta materiaalista 67 % kaikista kääntämättömistä sanoista, 78 % kaikista väärinkirjoitetuista sanoista, 57 % epäidiomaattisista lauseista ja 48 % kömpelöistä lauseista. Baskervillen koira saattoi kuitenkin olla myös huono materiaalivalinta, koska se koostuu jokseenkin vanhentuneesta kielestä, joka oli arvatenkin vielä vieraampaa kääntimille kuin moderni kaunokirjallinen teksti. Olisikin mielenkiintoista tutkia korpusta, joka koostuisi tuoreemmista teksteistä.

EU-korpusten tulokset olivat pääasiassa samankaltaisia, mutta viralliset kirjeet sisälsivät enemmän kielioppiin, kömpelyyteen ja epäidiomaattisuuteen liittyviä virheitä. Tämä tukee Biberin dimensioluokittelua, joiden mukaan viralliset dokumentit ovat informatiivisempia ja abstraktimpia kuin viralliset kirjeet. Virallisten dokumenttien keskimääräinen lausepituus oli myös pidempi (31,903 sanaa) kuin kahdella muulla korpuksella (viralliset kirjeet: 24,877, fiktio: 22,652), ja lausepituus havaittiin teorialuvussa vaikuttavan mahdollisesti neuroverkkokääntimien käännöslaatuun. Virallisten kirjeiden käännökset taas sisälsivät eniten poistoja (*omissions*), mutta tälle ei löytynyt mitään selvää syytä.




*Kaavio 1: Lingsoftin ja Googlen konekäänninten virheluokkien jakauma*

Käänninten välillä sekä Googlen että Lingsoftin neuroverkkokääntimet tekivät suurin piirtein samanlaisia virheitä, niin kuin kaaviosta 1 näkee. Ainoat merkitsevät erot olivat poistojen, pilkutuksen ja termivirheiden määrissä. Lingsoft teki enemmän poistoihin liittyviä virheitä, mutta toisaalta 48 % näistä oli fiktiokorpuksessa, mikä tukee ajatusta siitä, että kaunokirjallinen materiaali oli kääntimelle vierasta ja että käännintä oli ylipäätään koulutettu pienemmällä määrällä materiaalia. Googlen käännin taas teki enemmän pilkutukseen liittyviä virheitä, mikä johtuu kahdesta syystä: Googlen käännin lisäsi ylimääräisiä välilyöntejä ennen ja jälkeen tiettyjä merkkejä (/, <, >), mikä paremman kategorian puutteessa merkittiin pilkutusvirheeksi (*punctuation*). Googlen käännin oli myös toisinaan ilmeisen kykenemätön tuottamaan liian pitkien lauseiden käännöksiä loppuun asti: käännin keskeytti lauseita arbitraarisesti pisteellä ja syötti jäljelle jääneen materiaalin sen jälkeen välittämättä idiomaattisuudesta tai edeltävästä lauseesta (”[…] ja osallistumisen komiteoihin, ryhmiin ja työhön. osapuolten ja asiakirjojen valmisteluun”). Googlen käännös teki enemmän termivirheitä kuin Lingsoftin – varsinkin virallisten dokumenttien kanssa – mitä osaltaan voi jälleen selittää koulutusmateriaalin määrä. Jos Lingsoftin käännintä on koulutettu pienemmällä materiaalimäärällä, osaa se todennäköisimmin käyttää juuri EU-kontekstiin liittyviä termejä EU-käännöksissä. Jos Googlen käännintä taas on koulutettu valtavalla määrällä muutakin materiaalia, saattaa se EU-kontekstissa myös käyttää jonkin väärän kontekstin termiä.

Segmenttitason analyysin aikana ongelmaksi nousi eron tekeminen luokkien *mistranslation* ja *term error* välillä. Fiktio-korpuksen analyysiin ei esimerkiksi merkitty yhtäkään termivirhettä, vaikka joitain harmaalle alueelle sijoittuvia tapauksia olikin. Onko kyseessä esimerkiksi termivirhe, jos ”intolerant

(eyes)” kääntyy intoleranssiksi tai ”clutch (in which it held us)” auton kytkimeksi? Tutkimuksessa käytettiin TAUS:in (2019) määritelmää termistä (”domain-specific word”), mutta on kyseenalaista, voidaanko kaunokirjallisesta tekstistä ylipäätään löytää ainoastaan sille ominaisia sanoja. Eri asia tietysti on, kannattaako tällaista kategoriaa yleensä käyttää, jos sitä ei pysty soveltamaan kaikkien genrejen osalta tai jos se on vähintään ongelmallista.

### 4.3 Lausepituusanalyysi

Lausepituusanalyysin tulokset näyttävät, että lausepituudella ei voi luotettavasti ennustaa tekstien välisiä tuloksia: lyhyimpiä lauseita sisältänyt fiktiokorpus sai paljon huonomman kokonaistuloksen kuin pisimpiä lauseita sisältänyt viralliset dokumentit -korpus. Yhden korpuksen sisällä lausepituus sen sijaan näyttäisi jossain määrin korreloivan evaluaatiotuloksen kanssa ja lyhyemmät lauseet näyttäisivät ennustavan parempaa evaluaatiotulosta, joskin liian lyhyet lauseet olivat ainakin fiktiokorpuksessa kaikista huonoimpia tuloksia saaneita. Analyysin perusteella laatu näyttäisi kasvaa lauseiden lyhentyessä ainakin kuuteen sanaan asti, mutta esimerkiksi alle kolmen sanan lauseiden kääntäminen oikein näyttäisi olevan hyvin haasteellisia. Tälle on ymmärrettävä syy: mitä vähemmän lauseessa on sanoja, sitä todennäköisemmin käännin saa joko paljon oikein tai paljon väärin.

## 5 Lopputulokset

Tämä pro gradu -tutkielma on tarkastellut kielellisiä piirteitä, joita voisi käyttää arvioimaan ennalta erityyppisten tekstien soveltuvuutta neuroverkkokonekääntämiselle. Tutkielman perusteella kielellisiä piirteitä on mahdollista käyttää tällaiseen arviointiin, ja tutkielmassa on esitetty joitain alustavia tuloksia. Ei-narratiiviset tekstit näyttäisivät olevan soveltuvampia neuroverkkokääntämiselle kuin narratiiviset ja eksplisiittiset tekstit soveltuvampia kuin tilannekohtaiset. Myös lausepituutta voidaan käyttää jossain määrin ennustamaan tekstien soveltuvuutta, mutta luotettavampia tuloksia se antaa yhden tekstikorpuksen sisällä kuin useampien välillä.

Tutkimuksessa kritisoitiin myös nykyisten kääntimien ja analyysimenetelmien soveltuvuutta kaunokirjallisten materiaalien kääntämiseen ja arviointiin, ja arveltiin, että ne saattavat suosia merkittävästi niille tyypillisempiä tekstejä, kuten tutkimuksessa käytettyjä EU-korpuksia. Niin kuin teorialuvussa ilmaistiin, neuroverkkokääntimet toimivat parhaiten olosuhteissa, jotka simuloivat niiden koulutusta. Tutkimukseen valittu materiaali ja kääntimet saattoivatkin vääristää tuloksia EU-korpusten

suuntaan, ja olisi mielenkiintoista tehdä vastaava tutkimus hieman erilaisella materiaalilla ja eri tavalla koulutetuilla kääntimillä.

Tutkimuksen analyysin perusteella esitettiin, että neuroverkkokäänninten koulutusmateriaalin määrällä olisi merkittävä vaikutus käännöstulokseen ja että yksittäisen kielen ominaispiirteiden tunnistaminen ei olisi niiden kanssa välttämättä enää niin tärkeää. Toisaalta liiallinen koulutusmateriaalin määrä saattoi tehdä Googlen kääntimestä huonomman tunnistamaan juuri EU-kontekstissa esiintyviä termejä. Tätä ei kuitenkaan pysty lopullisesti todistamaan ilman tarkempaa tutkimusta.

Tutkielmassa esitetyt menetelmät olivat pääasiassa toimivia, mutta segmenttikohtaisen analyysin perusteella virhekategorisointiin esitetään seuraavaa muutosta. Virheluokat *mistranslation* ja *term errors* tulisi korvata selvemmällä kolmiportaisella erottelulla: 1) vääriin kieliopillisiin suhteisiin perustuva virhe, 2) väärä ala tai konteksti ja 3) väärä käännös ilman ilmeistä syytä. Tämä tekisi väärien käännösten luokittelusta selvempää ja tarjoaisi puitteet myös ei-tyypillisten tekstien arviointiin.

Neuroverkkokonekääntämisen soveltuvuuden tutkiminen on tutkimuksen perusteella monisyinen ongelma ja on selvää, että mikään yksittäinen palanen ei riitä ennustamaan tietyn tekstin soveltuvuutta. Vaikka lausepituus korreloikin tietyn korpuksen sisällä evaluaatiotulosten kanssa, ei sillä esimerkiksi voinut ennustaa tekstien välisiä eroavaisuuksia. Biberin dimensiot vaikuttivat lupaavilta, mutta vain kolmella genrellä tehty tutkimus antaa luonnollisesti ainoastaan alustavia tuloksia. Tutkimusta tulisikin jatkaa suuremmalla aineistomäärällä ja tässä tutkielmassa esitetyt menetelmät ovat siihen hyvä ensimmäinen askel.